THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Role of DNA supercoiling in genome structure and regulation

Samuel Corless

Ph.D

The University of Edinburgh

2013

Declaration of originality

I declare this thesis was written by me and is my own work.

Samuel Corless

Acknowledgements

I would like to thank Nick Gilbert for the opportunity to undertake my PhD in his lab and for his passionate supervision throughout the project. The many hours we have spent discussing my data and the wider chromatin field have been central to my development as a scientist and to the success of this project.

I would also like to thank the other members of the lab, who have each contributed through help, advice, tea and cake. Thank you to Catherine, Jayne, Ben, Jacqueline, Bernie, Ryu-Suke, Lora and Maria. I am also grateful to Adam Buckle for discussions throughout my project, to Shelagh Boyle for help with microscopy, to Duncan Sproul for bioinformatic assistance and to numerous members of the CGE section for reagents and advice. I would like to thank our collaborators for sharing resources and performing experiments that contributed to this work, in particular Mark Bradley's group in the Chemistry department and Bauke Ylstra's group at the VUMC microarray facility.

I am grateful to the Medical Research Council for funding my project and to the Institute of Genetics and Molecular Medicine for awarding me the PhD studentship. The IGMM has been a fantastic place to study and I am thankful for the colleagues that have made my time here so enjoyable, and for the many lasting friendships I have made.

I would like to thank my friends and housemates who have been a great support, you know who you are. In particular, I would like to thank Debbie Bradnock for proof reading this thesis.

Finally I would like to thank my family, whose support and encouragement made this possible.

Acknowledgement of collaborative work

Modern molecular biology is highly collaborative and I would like to acknowledge the contribution of Catherine Naughton, Nick Gilbert, Nicolaos Avlonitis, Christine Mordstein and the VUMC microarray facility to the experimental work contained within this thesis, as well as clarify my own contribution.

In Chapter 3 I performed all topoisomerase ChIP experiments and optimisations. The DNA was labelled and hybridised to microarrays by the VUMC microarray facility. I performed all subsequent analysis on the topoisomerase ChIP data, including comparisons with published RNA polymerase and DNA supercoiling datasets.

In Chapter 4 HPLC-MS was performed by Nicolaos Avlonitis. I performed all other characterisations of the bTMP molecule, including preparing the bTMP-DNA samples for HPLC-MS. The bTMP pull-down samples prepared by Catherine Naughton for Naughton *et al.* (2013) were labelled and hybridised to microarrays by the VUMC microarray facility. I performed all subsequent bioinformatic analysis.

In Chapter 5 I performed the characterisations of CFSs, with the exception of the quantification of the RPE1 CFS karyotype and the FISH image of the FRA3B fragile site which were performed by the project student Christine Mordstein. I optimised and performed the bTMP pull-down experiments, which were labelled and hybridised to microarrays by the VUMC microarray facility. I performed all subsequent bioinformatic analysis.

Abstract

A principle challenge of modern biology is to understand how the human genome is organised and regulated within a nucleus. The field of chromatin biology has made significant progress in characterising how protein and DNA modifications reflect transcription and replication state. Recently our lab has shown that the human genome is organised into large domains of altered DNA helical twist, called DNA supercoiling domains, similar to the regulatory domains observed in prokaryotes. In my PhD I have analysed how the maintenance and distribution of DNA supercoiling relates to biological function in human cells.

DNA supercoiling domains are set up and maintained by the balanced activity of RNA transcription and topoisomerase enzymes. RNA polymerase twists the DNA, over-winding in front of the polymerase and under-winding behind. In contrast topoisomerases relieve supercoiling from the genome by introducing transient nicks (topoisomerase I) or double strand breaks (topoisomerase II) into the double helix. Topoisomerase activity is critical for cell viability, but the distribution of topoisomerase I, IIα and IIβ in the human genome is not known. Using a chromatin immunoprecipitation (ChIP) approach I have shown that topoisomerases are enriched in large chromosomal domains, with distinct topoisomerase I and topoisomerase II domains. Topoisomerase I is correlated with RNA polymerase II, genes and under-wound DNA, whereas topoisomerase IIα and IIβ are associated with each other and over-wound DNA. This indicates that different topoisomerase proteins operate in distinct regions of the genome and can be independently regulated depending on the genomic environment.

Transcriptional regulation by DNA supercoiling is believed to occur through changes in gene promoter structure. To investigate DNA supercoiling my lab has developed biotinylated trimethylpsoralen (bTMP) as a DNA structure probe, which preferentially intercalates into under-wound DNA. Using bTMP in conjunction with microarrays my lab identified a transcription and topoisomerase dependent peak of under-wound DNA in a meta-analysis of several hundred genes (Naughton et al. (2013)). In a similar analysis, Kouzine et al. (2013) identified an under-wound

promoter structure and proposed a model of topoisomerase distribution for the regulation of promoter DNA supercoiling. To better understand the role of supercoiling and topoisomerases at gene promoters, a much larger-scale analysis of these factors was required. I have analysed the distribution of bTMP at promoters genome wide, confirming a transcription and expression dependent distribution of DNA supercoils. DNA supercoiling is distinct at CpG island and non-CpG island promoters, and I present a model in which over-wound DNA limits transcription from both CpG island promoters and repressed genes. In addition, I have mapped by ChIP topoisomerase I and IIβ at gene promoters on chromosome 11 and identified a different distribution to that proposed by Kouzine et al. (2013), with topoisomerase I maintaining DNA supercoiling at highly expressed genes. This study provides the first comprehensive analysis of DNA supercoiling at promoters and identifies the relationship between supercoiling, topoisomerase distribution and gene expression.

In addition to regulating transcription, DNA supercoiling and topoisomerases are important for genome stability. Several studies have suggested a link between DNA supercoiling and instability at common fragile sites (CFSs), which are normal structures in the genome that frequently break under replication stress and cancer. bTMP was used to measure DNA supercoiling across FRA3B and FRA16D CFSs, identifying a transition to a more over-wound DNA structure under conditions that induce chromosome fragility at these regions. Furthermore, topoisomerase I, IIα and IIβ showed a pronounced depletion in the vicinity of the FRA3B and FRA16D CFSs. This provides the first experimental evidence of a role for DNA supercoiling in fragile site formation.

Contents

8. Papers

Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nat. Struct. Mol. Biol. *20*, 387–395. Naughton, C., Avlonitis, N., **Corless, S.** *et al.* (2013).

Polymers for the cell-specific immobilisation of megakaryocytic cell lines. Macromol. Biosci. 13(4), 437-443. Hansen, A., **Corless, S.** *et al.* (2013).

Divergent RNA transcription: A role in promoter unwinding? Transcription *4*. Naughton, C., **Corless, S.**, and Gilbert, N. (2013).

List of abbreviations

°C	degrees Celsius
µg	$1x10^{-6}$ grams
µl	$1x10^{-6}$ litres
µM	$1x10^{-6}$ moles
A	adenine
A_{260}	absorbance at 260nm
bp	base pair
bTMP	biotinylated 4,5',8-trimethylpsoralen
C	cytosine
ChIP	chromatin immunoprecipitation
ChIP-chip	ChIP hybridised to microarrays
ChIP-seq	ChIP sequencing
CFS	common fragile site
CpG	cytosine guanine dinucleotide in which cytosine is 5'
cm	$1x10^{-2}$ metres
Cy3	cyanine dye that fluoresces yellow-green
Cy5	cyanine dye that fluoresces red
Dam-ID	DNA adenine methyltransferases identification
FISH	Fluorescent *in situ* hybridisation
g	grams
G	guanine
IF	immunofluorescence
kb	$1x10^{3}$ base pairs
KDa	$1x10^{3}$ Daltons
L	litres
log2	binary logarithm

M	moles
Mb	$1x10^{6}$ base pairs
mg	$1x10^{-3}$ grams
min	minutes
mm	$1x10^{-3}$ metres
mM	$1x10^{-3}$ moles
mg	$1x10^{-3}$ grams
ml	$1x10^{-3}$ litres
MW	molecular weight
nm	$1x10^{-9}$ metres
rpm	revolutions per minutes
SDS	sodium deodecyl sulphate
T	thymine
TSS	transcription start site
U	units
UV	ultra-violet
V	volts

List of Figures

List of tables

1. Introduction

The principle challenge of the post-genomic era is to understand how a human genome is regulated within a cell, and how the coordinated regulation of genetic information in each cell contributes to tissue, organ and organism structure and function in health and disease. The field of chromatin biology has documented a variety of proteins, protein modifications and DNA modifications that reflect and determine the transcription and replication state of a genomic region. The ENCODE project attempts to take this functional classification further by mapping many of these factors at high resolution across several defined cell types (Dunham et al., 2012). However, this catalogue of bound proteins and chemical modifications only covers part of the regulatory potential within chromatin.

DNA molecules are not uniform linear sequences of nucleotide pairs, but varied and dynamic structures that contribute to their own packaging and regulation. Alternative DNA helices, alternative DNA structures and changes in the helical twist of the DNA molecule are all likely to contribute to the regulation of the human genome (Bates and Maxwell, 2005). In addition, a number of diseases have been associated with alternative DNA structures, either directly (e.g. fragile X syndrome (Bacolla and Wells, 2009)) or through increased genome instability (e.g. common fragile sites (Zlotorynski et al., 2003)). Changes in DNA helical twist, called supercoiling, are of particular interest as they have been shown to regulate transcription *in vitro* and in prokaryotes (Peter et al., 2004; Tabuchi and Hirose, 1988, 1988; Weintraub et al., 1986). Our lab recently identified domain scale changes in DNA helical twist in the human genome, called supercoiling domains, which echo the regulatory domains of prokaryotes (Naughton et al., 2013a). As in prokaryotes, these DNA supercoiling domains are maintained by transcription and topoisomerase activity. However, the individual roles of topoisomerase I, IIα and IIβ in regulating DNA supercoiling in the human genome are unknown, as their genomic distributions have not been well characterised. In addition, although the relationships between topoisomerases, RNA polymerase II and DNA supercoiling are critical for

transcription at gene promoters (Lee et al., 1993; Lyu et al., 2006; Sano et al., 2008; Tabuchi and Hirose, 1988), the distribution of topoisomerases and DNA supercoiling at promoter regions have been poorly characterised (Kouzine et al., 2013; Naughton et al., 2013a). Understanding the distribution of topoisomerases and DNA supercoiling at different scales will help identify their relationship with genome regulation. Furthermore, DNA supercoiling is thought to influence genome stability and contribute to the expression of common fragile sites (CFSs) (Burrow et al., 2010; Gellibolian et al., 1997), but the distribution of DNA supercoils has not been determined at these regions *in vivo*. This thesis aims to characterise the inter-relationships between topoisomerases and DNA supercoiling and their joint role in the regulation of gene expression and genome stability in human cells.

1.1 Chromatin structure and regulation

Each human cell contains almost 2 metres of DNA coiled and folded into a nucleus 6 μM in diameter in a nucleoprotein structure called chromatin (Wolffe, 1998). This feat of compaction is performed at several levels, from the coiling of the DNA molecule to the condensation of entire metaphase chromosomes. The standard model of chromatin structure (Figure 1.1) posits the polymer of nucleotide pairs twists into a double helical conformation and winds 147 base pairs (bp) of DNA around histone protein octamers connected by short regions of 'linker' DNA. This forms a 'beads on a string' structure which coils into a 30 nm fibre, higher order-fibres and on to large scale chromatin structures. This nucleoprotein structure provides multiple scales at which genes can be regulated, although many of the mechanisms are poorly understood.

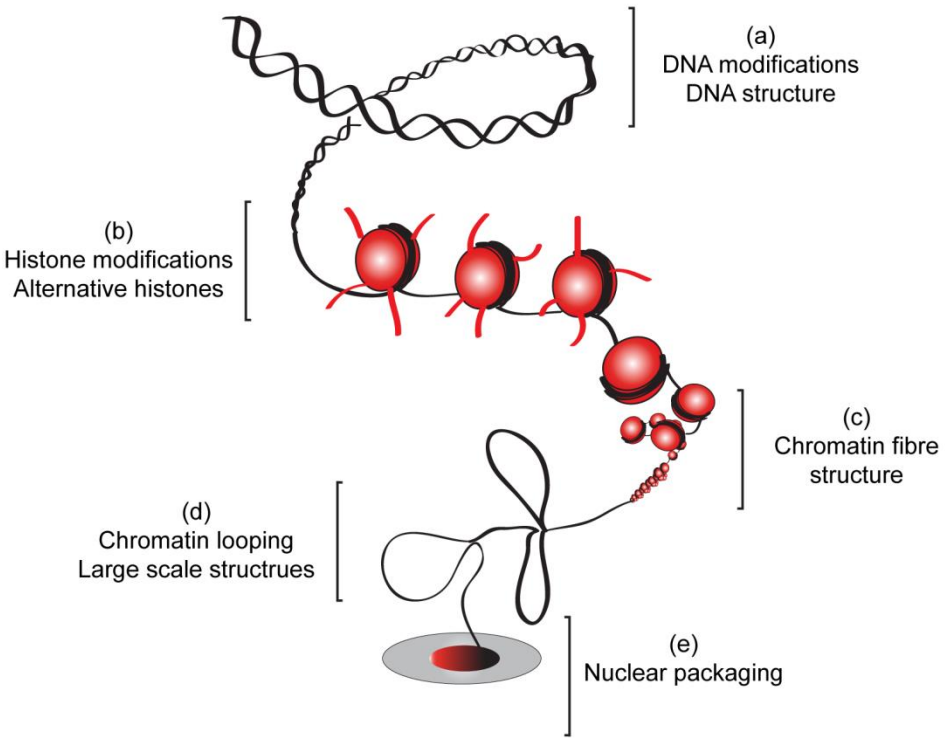


Figure 1.1 DNA packaging into chromatin. The DNA double helix (a) wraps around histone octamers and together with linker histones forms nucleosome arrays (b) which coil into higher order chromatin fibres (c) that further organise into large scale chromatin structures (d) and territories within the nucleus (e).

1.1.1 DNA

DNA is a polymer of deoxyribonucleotide pairs which carries the genetic information and forms the backbone of chromatin structure (Figure 1.1a). Although frequently represented as a linear sequence of base pairs onto which important regulatory proteins and modifications bind (e.g. UCSC genome browser), in reality the heterogeneous structure of individual bases and their combined structure in a DNA molecule carry significant regulatory potential.

1.1.1.1 Nucleotide biochemistry

The fundamental components of DNA are the nucleotide bases (Bates and Maxwell, 2005; Watson and Crick, 1953). The structure of each nucleotide can be partitioned into a phosphate group, deoxyribose sugar and a variable nucleotide base (Figure 1.2a). The phosphate and deoxyribose components constitute the backbone of DNA, forming phosphodiester linkages between consecutive bases on a single strand of the DNA helix. The variable nucleotide bases interact between complementary single stranded DNA molecules to form the double helix, through either adenine to thymine (A to T) or guanine to cytosine (G to C) interactions (Figure 1.2a). Together these nucleotide interactions make up the structure of the double helix. Vitally, the order with which the nucleotides are arranged within this double helix forms the genetic blue print of an organism.

The signature defined by the genetic code is based on differences in the chemical structure of each nucleotides variable base. Nucleotides can be separated on the structure of their variable base into purines (adenine and guanine) and pyrimidines (cytosine and thymine), based on the presence or absence of an imidazole ring (Figure 1.2a) (Bates and Maxwell, 2005). The distinct structure of each nucleotide exposes different chemical bonds and shapes for the interaction of DNA binding proteins and modification by DNA modifying enzymes. Furthermore, the structure and stability of base-pairing is influenced by the chemistry of nucleotide interactions, with GC base pairs having three hydrogen bonds that confer a stronger interaction compared to the two hydrogen bonds of AT nucleotides under the rules of Watson-

Crick base pairing (Figure 1.2a). In addition the propeller twist of a base pair, in which one base is twisted relative to the other in a similar way to the propeller of a helicopter, is different between AT and GC base pairs in a sequence context dependent manner and can vary between 5° and 25° (Calladine, 2004) (discussed further in Section 1.1.1.2).

In addition to variability in the structure of canonical nucleotides, chemical modifications can influence the chemical and structural properties of nucleotides and the DNA helix. The best studies example of DNA nucleotide modification is methylation of cytosine at position 5 (Figure 1.2b), with some studies referring to this as the 'fifth base' (e.g. Wu et al., 2010). This modification acts as a docking site for methyl-cytosine specific DNA binding proteins and promotes the formation of the Z-form alternative DNA helix (Section 1.1.1.4).

It is critical to understand this heterogeneity in nucleotide structure, base pair stability, propeller twist and chemical modification to appreciate genome structure and regulation at the level of the DNA helix and beyond.

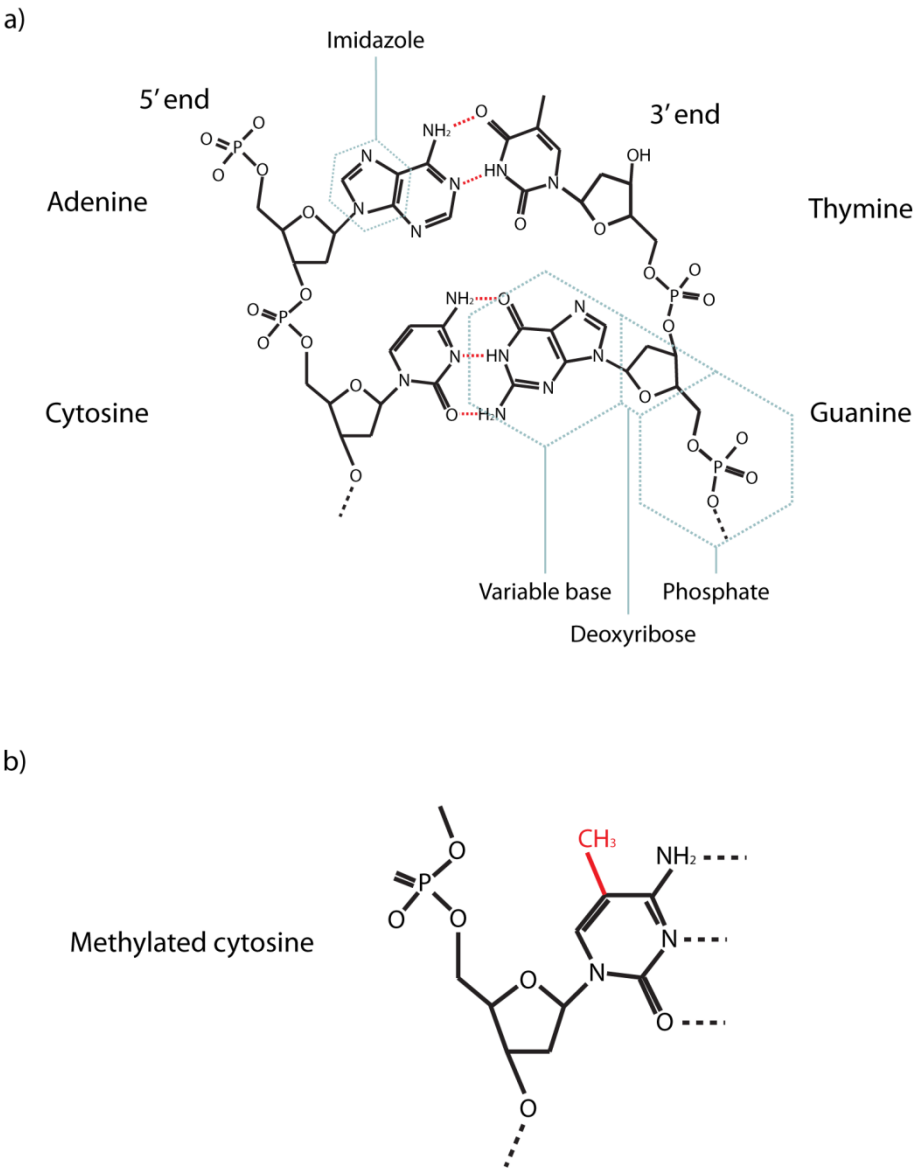


Figure 1.2 The chemical structure of DNA base pair interactions. a) The interactions between adjacent bases on the same strand are through phosphodiester bonds. The interactions between base pairs are through hydrogen bonds and are identified by a broken red line. General nucleotide components are identified by broken grey lines. b) Chemical structure of methylated cytosine with the modification shown in red.

1.1.1.2 Double helix structures

The discovery of the DNA double helix identified the structure and mechanism by which genetic information is stored, transcribed and replicated (Watson and Crick, 1953). The double helix is one solution to the challenge of packaging hydrophobic nucleotide bases within the water rich environment of the nucleus (Calladine, 2004). A simple ladder design for DNA, with the hydrophobic bases in the centre and the hydrophilic backbone either side, would expose bases to water and be unstable. However, a twisted ladder allows bases to stack on top of one another in a way that prevents their exposure to water. This twisted ladder is the double helix.

Helical structure can be profoundly influenced by the properties of nucleotides in base-pairs and within their local sequence context. A major structural property of individual base pairs is differences in propeller twist, as discussed in Section 1.1.1. Propeller twist influences flexibility at the dinucleotide step, with high propeller twist making stacking into a dinucleotide step more awkward (El Hassan and Calladine, 1996). Therefore, sequences that generate high propeller twist produce structural discontinuities in the DNA, which may be recognised by DNA binding proteins. In addition, structural properties of base pairs with respect to one another that can influence DNA helical structure, most notably differences in twist, roll and slide (Calladine, 2004). Twist is the rotation of one base pair relative to another, about an axis that runs vertically through neighbouring base pairs. This differs from roll, which is the rotation of one base pair relative to another along its long axis (i.e. one that runs horizontally through a base pair). Finally, slide is the position of a base pair relative to the adjacent base pair along its long axis. These three properties vary between different sequence combinations and can generate structural discontinuities and different double helical structures, including favouring alternative DNA helices.

A popular misconception is that the structure of DNA is a regular repeating pattern of the double helix identified by Watson and Crick (1953). This B-form DNA, with a deep major groove and shallow minor groove (Figure 1.3), may represent much of the DNA *in vivo* but there are other helices including A-form and Z-form DNA. Both A- and B- form DNA are right handed helices, but the major and minor groove of A-form DNA have similar width and depth (Figure 1.3) (Bates and Maxwell,

2005). This exposes a distinct charge distribution which may influence the binding of specific A-DNA binding proteins (Rohs et al., 2010). Importantly, almost all external features of A- and B- form DNA can be determined from their values of twist, roll and slide (Calladine, 2004). On the other hand, Z-form DNA represents a major structural rearrangement in which a left handed helix forms a 'zig-zag' pattern (Figure 1.3), which is almost 'inside out' compared to A- or B- form DNA (Wang et al., 1979). This exposes another charge distribution for the formation of DNA-protein interactions (Rohs et al., 2010). The structures of DNA bases that form Z-form DNA are not well understood (Calladine, 2004), but it is likely that a major factor in this structure is unusual values for twist, roll and slide. These 'alternative' DNA helices were originally thought to be effects produced by structural techniques, but substantial evidence now indicates that they are present in DNA and are potentially functional *in vivo*. Common sequence and context dependent regions of A form and AB intermediates have been identified by crystallography and NMR, indicating that similar structures can be seen under different experimental conditions (Hays et al., 2005; Ng et al., 2000). Furthermore, antibodies to Z-DNA structures have identified this structure in *Drosophila* and human (Nordheim et al., 1981; Wittig et al., 1991, 1992). In addition, all three DNA helices have been crystallised in association with DNA binding proteins, including TATA binding protein with A-form DNA (Kim et al., 1993; Lu et al., 2000) and ADAR1 with Z-form DNA (Schwartz et al., 2001). Therefore, at least three forms of DNA helix occur *in vivo*. Whether these proteins bind DNA helices in the alternative conformation, induce the conformation or stabilise a transient structure is currently unknown.

Understanding the distribution of different DNA helices *in vivo* may to help identify their function and the mechanisms that regulate their distribution. At present there are no techniques for mapping the relative distribution of A- and B- form DNA, although one potential method could be to use chemical probes that differentiate between the width and depth of the major and minor grooves. On the other hand, the distribution of Z-DNA has been mapped at a small number of locations using anti-Z DNA antibodies to regions of the fly (Nordheim et al., 1981) and human genome (Wittig et al., 1991, 1992). Z-DNA is generally identified at GC rich regions of the genome, particularly in the vicinity of gene promoters, and is present in a DNA

supercoil dependent manner(Wittig et al., 1989). The identification of ADAR1 as a specific Z-DNA binding protein, with an unknown function in the context of this alternative DNA helix, suggests a selective pressure for Z-DNA binding proteins (Herbert et al., 1995, 1997). Therefore, the relationship between transcription and DNA helical structure in the vicinity of gene promoters indicates an potential relationship between DNA structure and gene activity, but the mechanism and function of this relationship *in vivo* remains unknown.

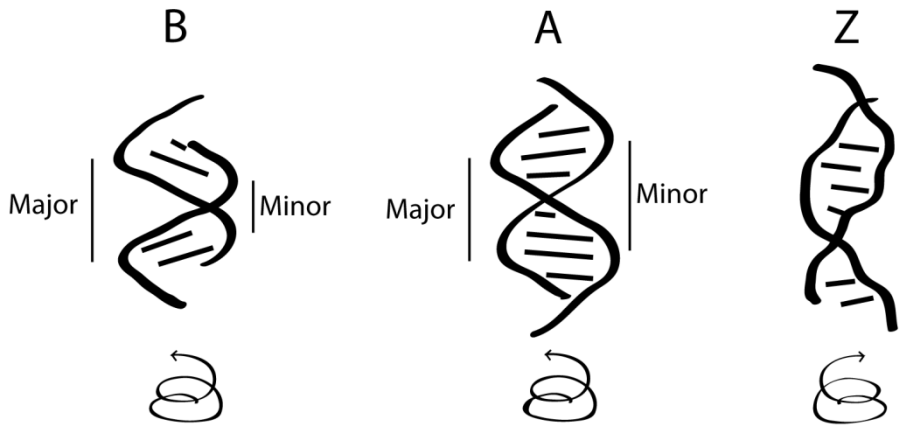


Figure 1.3 DNA helical structure. The three most common forms of the DNA double helix are A, B and Z form. B form is the typical form identified by Watson and Crick (1953) and is characterised as a right handed helix with a clear difference in groove width between the major and minor groove. The A form is also a right handed helix, although the major and minor grooves are similar in size. Z form DNA is a left handed helix with no distinct major or minor groove.

1.1.1.3 Alternative DNA structures

Heterogeneity of DNA structure is not limited to double helices, with other alternative DNA structures including cruciforms, triple helices and G-quadruplexes being described (Bates and Maxwell, 2005; Brázda et al., 2011; Lipps and Rhodes, 2009; Shlyakhtenko et al., 2000). Cruciform structures are formed from the unwinding and pairing of two single strands to form intra-strand double helices (Figure 1.4). They can occur on sequences that form palindromes in the double stranded DNA, where a sequence is followed immediately or with a short linker by a reverse of the same sequence (Bates and Maxwell, 2005). Despite unfavourable thermodynamics, there is substantial evidence that cruciform structures form *in vivo* and are important in the regulation of some genes (reviewed in Brázda et al., 2011). Similar to cruciforms, DNA triplexes form on under-wound DNA through an interaction of the double helix with a third single stranded DNA in a parallel or anti-parallel orientation (Figure 1.4) (Bates and Maxwell, 2005). In this structure the third DNA strand lies in the major groove, forming non-Watson-Crick base pairs with purine bases. This leaves a single stranded DNA that was originally paired with the third helix in the triplex. The presence of DNA triplexes is associated with some triplet repeats *in vivo* (Wojciechowska et al., 2006), which have been implicated in neurological conditions including fragile X syndrome (Section 1.3). Similarly, G-quadruplexes are a four-stranded structure formed by two hairpins in opposing DNA strands (Figure 1.4) (Bates and Maxwell, 2005). They are associated with genomic features including centromeres and telomeres, as well as important processes including transcription, recombination and replication (Biffi et al., 2013; Lipps and Rhodes, 2009). A recent study confirms the presence of these structure *in vivo*, identifying a cell-cycle dependent regulation consistent with a replication dependent formation *in vivo* (Biffi et al., 2013). A common feature of all alternative structures is the requirement for a stretch of nucleosome-free unwound DNA. Localised under-winding of DNA occurs in transcription dependent DNA supercoiling (Section 1.2.4), again supporting a relationship between DNA structure, transcription and gene regulation *in vivo*.

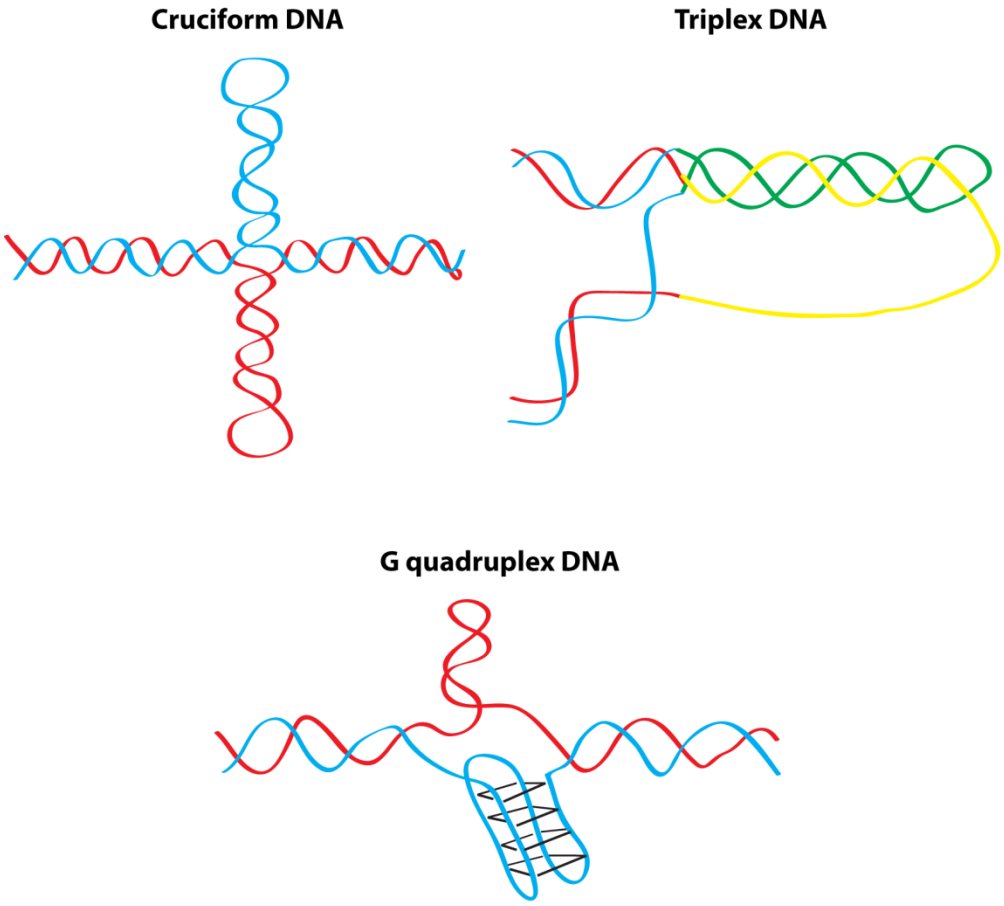


Figure 1.4 Alternative DNA structures. For cruciform and G-quadruplex structures one strand of the DNA is coloured red and the other blue. In the case of triplex DNA, the colouration is supplemented with green to represent Watson and Crick base-pairing in the triplex and yellow represents non-Watson and Crick base-pairing.

1.1.1.4 DNA sequence motifs

The order of nucleotides has a number of vital functions within the genome, from encoding proteins to regulating gene expression and genome structure. In general, proteins detect specific binding motifs by first recognising a large-scale feature of the molecule such as the shape/curvature of the helix and then probing the details of the bases within (Calladine, 2004). Therefore, important structural properties at sequence motifs include both DNA helical structure (Section 1.1.1.2) and nucleotide chemical structure (Section 1.1.1.1).

One example of a transcription factor associated with a DNA sequence motif is that TATA box binding protein (TBP) which binds the TATA-box nucleotide motif with a degree of sequence redundancy (Calladine, 2004). Structural studies of the interaction between TBP and the TATA box identify an A-form DNA helix within the protein binding site (Kim et al., 1993). Whether TBP identifies DNA in an A form conformation or regions that have a propensity to form A form DNA upon protein binding is unknown, but it may be this property of DNA structure that TBP initially identifies before probing the sequence of the bases within. Other transcription factors and structural proteins have consensus binding sequences, which are also likely to combine helical structure and nucleotide sequence to determine interactions *in vivo* (Portales-Casamar et al., 2010). These DNA sequence motifs represent the smallest scale of DNA structure with specific regulatory potential.

In addition to short motifs of specific sequence there are longer regions of general nucleotide enrichment which have regulatory potential. The best example is that of CpG islands, which are regions of elevated CpG density found at the promoter of ~70% of human genes (Ehrlich et al., 1982). CpG rich DNA sequences have been shown to form G-quadruplexes, Z-form DNA and other alternative structures (Section 1.1.1.3), particularly in the vicinity of promoter regions. Furthermore, methylation of the cytosine nucleotide (Section 1.1.1.1) at CpG island promoters correlates with gene repression (Keshet et al., 1985; Weber et al., 2007). This modification recruits methyl-CpG specific histone modification proteins (Jones, 2012) and stabilises Z-DNA helices (Behe and Felsenfeld, 1981) which may mediate gene repression, although no direct causal relationship has been determined. Other

regions elevated for nucleotide sequence at a large scale include the repeat regions of telomeres, centromeres and rare fragile sites (Gellibolian et al., 1997) as well as the AT-rich regions found at common fragile sites (Section 5.1). In each case, DNA helical structure has been implicated in stability and function at these regions of the genome. This indicates that large scale sequence distributions are important for the regulation of transcription, replication and genome stability *in vivo*.

1.1.2 Nucleosomes

The first level of packaging of eukaryotic DNA into chromatin is the nucleosome, made up of a core particle, linker histone and linker DNA (Figure 1.1b). The core particle is an octamer of histone proteins constraining 147 bp of DNA to its outer surface (Davey et al., 2002; Luger et al., 1997). Core particles are connected by linker DNA, which varies in length between 7 bp and 101 bp in an organism and cell line dependent manner (Van Holde, 1989). An array of nucleosomes form the classic 'beads on a string' structure (Thoma et al., 1979). Each of the constituents of this lower-order chromatin fibre can be exchanged with alternative proteins or chemically modified, influencing chromatin structure and function.

In the context of the nucleosome the DNA molecule is frequently portrayed as a fibre wrapped around a cylinder, much like cotton wrapped around a reel. It is important to note that the structure of the nucleosome is influenced as much by DNA sequence/structure as by protein structure (Rohs et al., 2010). In this equal partnership, the position of the core particle on the DNA backbone, the length and structure of the linker DNA and the subsequent folding into higher order chromatin fibres will each be influenced by heterogeneity within the DNA molecule (Section 1.1)

1.1.2.1 Core histones

Core histone proteins compact the DNA and make protein-protein interactions with themselves and other chromatin binding proteins (Wolffe, 1998). The core particle

comprises pairs of the histone proteins H2A, H2B, H3 and H4 which bind 147 bp of DNA to the outer surface through electrostatic interactions (Yager et al., 1989). Each core histone is a small basic protein (11-16 KDa) with relatively high levels of argenine and lysine (>20% of amino acids). Their structure comprises a histone-fold domain at the carboxyl (C-) terminus that interacts with DNA and other core histones (histone core), and a charged amino (N-) terminus that contains the majority of lysine residues (histone tail) (Arents et al., 1991; Wolffe, 1998). Post-translational modifications of the N-terminal tails or replacement of core histones with specific histone variants can indicate altered states of genome regulation, including transcription and DNA repair.

1.1.2.2 Histone modifications

Post-translational modifications of histone proteins generally reflect the underlying regulation of the genome. For example trimethylation of histone H3 lysine 4 (H3K4me3) indicates an active promoter, whilst trimethylation of histone H3 lysine 27 (H3K27me3) indicates transcriptional inactivity (Ernst et al., 2011; Kouzarides, 2007). Using a combination of histone modification distributions, defined by ChIP-seq, several groups have attempted to identify functional classes of chromatin *in vivo* (Ernst et al., 2011; Hon et al., 2009). One classification mapped nine histone modifications and several chromatin associated proteins, establishing 15 chromatin states which include 'active promoter', 'strong enhancer', 'weak transcribed' and 'polycomb repressed' (Ernst et al., 2011).

The identification of chromatin states provides a useful segregation of the genome, but in most cases the relationship between histone modifications and genome regulation is purely correlative. There are two mechanisms by which histone modifications have been shown to directly alter chromatin structure or gene regulation, through the disruption of nucleosome-nucleosome contacts and the recruitment of non-histone protein. The stability of fibre-fibre interactions can be influence by histone acetylation, which neutralises the basic charge of lysines on the histone tail (Kouzarides, 2007; Wolffe, 1998). For example, acetylation of H4K16 has been shown to disrupt the formation of the 30 nm fibre (Shogren-Knaak et al.,

2006). The relationship between fibre-fibre interactions and other histone modifications is largely unknown, although it has been suggested that phosphorylations may also influence nucleosome interactions (Kouzarides, 2007). The recruitment of non-histone proteins by histone modifications is better characterised, with chromodomain proteins recognising methylation, bromodomain proteins recognising acetylation and 14-3-3 proteins recognising phosphorylations (Kouzarides, 2007). Examples of proteins which specifically interact with particular histone modifications include JMJD2A, CHD1 and NURF to H4K4me, PC2 to H3K27me and HP1 to H3K9me (Huang et al., 2006; Kouzarides, 2007; Sims et al., 2005; Wysocka et al., 2006).

To identify if specific histone modifications are essential for genome regulation *in vivo* several groups have knocked-out individual modifying enzymes in cell lines and model organisms. Knock-outs for methyltransferases that modify H3K9me1 and H4K20me1 are essential for mouse development and survival, but not cell survival or replication (Dodge et al., 2004; Pinheiro et al., 2012). On the other hand, knocking-out the methyltransferase that modifies H3K9me3 is non-lethal, although the mice survive at sub-Mendelian frequencies and have a predisposition to cancer (Peters et al., 2001). In each of these knock-out experiments a significant increase in genome instability was observed, although this did not prevent cell survival or replication, indicating a context or cell-type dependent requirement for specific modifications *in vivo*. How other histone modifications influence genetic regulation, cell survival and development is difficult to establish through knock-out experiments, as one protein can often modify multiple sites and one site can be modified by multiple proteins. For example, the CBP/300 protein can acetylate at least seven different lysine residues on four different histones, whilst the H3K4me3 modification can be placed by at least nine different proteins (Kouzarides, 2007). In addition, many histone-modifiying enzymes have additional non-histone substrates and It may be that knock-out experiments influence cell survival and genome stability through indirect mechanisms. Finally, our understanding of histone modifications is further limited by current research trends, which are focused on methylation and acetylation whilst neglecting the vast repertoire of phosphorylations, deaminations, ADP-ribosylations, ubiquitinations and sumoylations (Bannister and Kouzarides, 2011). Therefore,

although histone modifications are assumed to have biological consequences, in many cases these are unproven in the context of the local chromatin state.

1.1.2.3 Histone variants

Canonical histone proteins within a core particle can be replaced with variant histones for H2A, H2B and H3. These histone variants reflect structural and regulatory specialisations within the chromatin structure (Sarma and Reinberg, 2005). For example, the histone H3 variant CENPA is found exclusively in centromeres where it performs an essential function, with CENPA knock-out mice exhibiting early embryo lethality (Howman et al., 2000). Other histone variants identify regions of active transcription (H2ABBD), enhancer/promoters (H2AZ) or X inactivation (macroH2A) (Costanzi and Pehrson, 1998; Ku et al., 2012; Tolstorukov et al., 2012). The post-translational modification of histone variants can further indicate genomic changes, with the phosphorylation of serine 139 of H2AX representing an early step in the repair of double strand DNA breaks (Rogakou et al., 1998). Therefore, histone variants and their modifications present a further indication of genetic regulation at the scale of the nucleosome.

1.1.2.5 Linker regions

Linker DNA varies from 7 bp to 101 bp in an organism and cell type specific manner and is usually bound by the linker histone protein histone H1 (Van Holde, 1989). The linker histones which bind the DNA exiting the core particle (Thomas, 1999) have been shown *in vitro* to stabilise the chromatin fibre (Thoma and Koller, 1977; Thoma et al., 1979), inhibit transcription (O'Neill et al., 1995) and prevent nucleosome sliding (Pennings et al., 1994). The linker DNA is of particular significance for the study of DNA structure, as this nucleosome free DNA can form alternative DNA structures and contain unrestrained DNA supercoils in the context of chromatin (Branello et al., 2012).

1.1.2.6 Nucleosome positioning

There is a dynamic competition for DNA binding between transcription factors and nucleosomes, the balance of which can be regulated to influence transcription at gene promoters. For example, constitutively expressed genes have an 'open' promoter structure which is often depleted for nucleosomes over a region ~150 bp upstream of a transcription start site, and is therefore available for transcription factor binding (Cairns, 2009). On the other hand, regulated genes have 'covered' promoters that position nucleosomes over the promoter region to limit transcription factor binding and rely on an active chromatin remodelling complex to reposition nucleosomes and activate genes (Cairns, 2009). A number of sequence parameters regulate the DNA positioning on the surface of the core particle, based primarily on the variable path of DNA around the histone octamer. In this path the first and last 10 bp of DNA are almost straight (Richmond and Davey, 2003), followed by curved DNA until two very sharp bends approximately one and four helical turns either side of the centre of the core particle DNA (Hogan et al., 1987; Wolffe, 1998). Owing to the fact that some DNA sequences are intrinsically straight, curved or easily bent, there is a degree of sequence preference for the formation of a core particle. Important features include the overall ability of the 147 bp to bend around a histone octamer (Drew and Travers, 1985; Struhl and Segal, 2013) and the deformability of the DNA 20-30 bp either side of the DNA centre (FitzGerald and Simpson, 1985; Wolffe, 1998). The stability of DNA interactions with the surface of the core-particle can be further enhanced by a 10 bp periodicity of A-tracts which form a narrow minor groove on the inside edge of the DNA helix, bending the DNA around the core histones (Rohs et al., 2009). In yeast, the periodicity of A-tracts is a general feature of genome organisation (Segal et al., 2006) and it is predicted that similar structural features occur across eukaryotes (Struhl and Segal, 2013). To test the influence of sequence on core-particle positioning *in vivo* Segal et al. (2006) developed a tool to predict core-particle position in the yeast genome. This tool predicts the position of ~50% of nucleosomes to within 35 bp, indicating that DNA sequence is one determinant of core histone position *in vivo*. In this manner, DNA sequence and structure influences nucleosome position and therefore regulatory potential. For example, the primary determinant of a 'open' chromatin structure is an underlying

enrichment for poly(dA:dT) sequences that disfavour nucleosome binding (Hughes et al., 2012; Segal and Widom, 2009), a feature absent from 'covered' promoters. A sequence common at 'covered' promoters is the TATA box, which preferentially binds the surface of the core-particle preventing TBP binding and subsequent transcription (Patikoglou and Burley, 1997). Activation of 'covered' promoters requires the switching, sliding or removal of nucleosomes from transcription factor binding sites by chromatin remodelling complexes.

Complementing the sequence based positioning of nucleosomes, chromatin remodelling complexes restructure, mobilise and eject nucleosomes in an ATP-dependent manner to regulate access to DNA (Saha et al., 2006). For example, there is a strong positioning of the +1 nucleosome upstream of the TSS in yeast that is dependent on chromatin remodellers but not on sequence (Zhang et al., 2011). There are five known families of chromatin remodellers that affect the structure of nucleosome and nucleosomes arrays in distinct manners; SWI/SNF, ISWI, NURD/Mi-2/CHD, INO80 and SWR1. SWI/SNF and ISWI have been studied in particular detail and have very different functions *in vivo*. ISWI is important for the ordering and phasing of nucleosomes on the chromatin fibre following replication, with a role in promoting gene activation and repression (Corona and Tamkun, 2004). SWI/SNF, on the other hand, disorders nucleosomes at promoter regions, which can promote or disrupt the binding of transcription factors and gene activation (Martens and Winston, 2003; Saha et al., 2006). The mechanism of repositioning by 'nucleosome sliding', shared by SWI/SNF and ISWI, involves binding a nucleosome at a defined position and breaking the histone-DNA interactions to create a wave of DNA tension (i.e. DNA supercoiling) on the octamer surface that propagates around the nucleosome in a direction determined by the chromatin remodelling complex (Saha et al., 2006). Whether unconstrained DNA supercoiling can similarly influence nucleosome positioning *in vivo* is uncertain, although a recent meta-analysis of 445 transcription start sites suggests that nucleosome positioning is not significantly affected by differences in unrestrained supercoiling (Kouzine et al., 2013). However, a minority of promoters with sequence features that have the propensity to form alternative DNA helices or structures may strongly influence nucleosome positioning. In this way, DNA supercoiling could contribute to the

remodelling of nucleosomes with and/or without the aid of chromatin remodelling complexes.

1.1.3 Higher-order chromatin fibres

The lower order chromatin fibre represents the first level of chromatin compaction, with nucleosomes compacting the genome ~7 fold over naked DNA (Németh and Längst, 2004). In order to fit the human genome within a cell nucleus the genome must be compacted ~10,000 fold (Calladine, 2004), through a hierarchy of higher-order chromatin fibres (Figure 1.1c). Visualising the compaction of chromatin fibres under different ionic conditions *in vitro* indicates that the first level of higher order compaction is a 30 nm fibre (Thoma et al., 1979). This 30 nm fibre is believed to form a solenoid (Kruithof et al., 2009) or zigzag (Schalch et al., 2005) helical structure, which compacts the genome by ~50 fold (Németh and Längst, 2004). Despite extensive evidence that chromatin folds into the 30 nm fibre and beyond under physiological conditions (reviewed in Wolffe (1998)), the presence of a 30 nm fibre has not been conclusively demonstrated *in vivo* and remains controversial (Fussner et al., 2011; Maeshima et al., 2010; Staynov, 2008). Beyond the 30 nm fibre compaction of the chromatin fibre is even less well defined, with additional folding and coiling predicted to form a ~100 nm chromonema fibres, 200-300 nm fibres and eventually metaphase chromosomes (Bak et al., 1977; Belmont and Bruce, 1994; Sedat and Manuelidis, 1978; Taniguchi and Takayama, 1986).

1.1.4 Large-scale chromatin structures

The chromatin fibre has been shown to form large-scale looped domains by both cytological and molecular biology techniques. Large-scale DNA loops were first identified by electron microscopy of *E. coli* chromosomes in which loops averaging 38 to 77 kb were seen emanating from a central scaffold in a rosette like structure (Kavenoff and Ryder, 1976). The separation of the *E. coli* genome into distinct supercoil domains is further supported by a molecular study that determined the number of DNA nicks required to relax all supercoils in the genome (Worcel and

Burgi, 1972). In this study DNA is released in a highly folded conformation by careful lysis of *E. coli* cells and analysed by sucrose gradient sedimentation following treatment with DNaseI. DNaseI introduces nicks into the DNA that relieve supercoils from the genome. In *E .coli* supercoils compact the genome through writhe (Section 1.2.2), therefore nicking leads to a less compact genome that sediments more slowly through a sucrose gradient. By varying the concentration of DNaseI, the authors establish that between 6 and 40 nicks are required to relieve all DNA supercoils. These results are interpreted as a separation of the *E. coli* genome into between 12 and 80 distinct DNA loops with an average size of between 58kb and 383kb. In eukaryotes a similar segregation of the genome into distinct DNA supercoil domains was identified in the interphase genome of *Drosophila* using a similar nicking technique, which identified distinct domains of ~85 kb (Benyajati and Worcel, 1976). In addition, looped structures have been observed cytologically in human metaphase chromosomes, although it is not known whether these are preserved in interphase (Earnshaw and Heck, 1985; Paulson and Laemmli, 1977). Together this data identifies that prokaryotic and eukaryotic genomes are organised into distinct 50-100 kb domains which may be related to cytologically defined DNA loop structures. Therefore, the compartmentalisation of the genome into looped domains may represent a conserved mechanism for the separation of structure and function.

To establish the distribution of looping in the interphase genomes at high resolution and in three dimensions the 'chromosome conformation capture' (3C) method was developed (Dekker et al., 2002). In this technique nuclei are cross-linked using formaldehyde so that regions of the genome that are in physical contact are bound covalently to neighbouring proteins and DNA. Cross-linked DNA is digested with a restriction enzyme and DNA strands are ligated at a very low DNA concentration, so that the ligation of cross-linked fragments is much more probable than between random fragments. To identify cross-links and quantify their frequency, ligation products are analysed by PCR or deep sequencing. Using variations on the 3C technique (reviewed in De Wit and De Laat, 2012), interactions have been identified at scales from kilobases to megabases. For example, the transcription dependent interaction of the HoxD gene with its long range regulatory elements (Montavon et

al., 2011), and all DNA interactions within a 4.5 Mb region surrounding the X-inactivation centre (Nora et al., 2012), indicating a relationship between genome structure and gene regulation *in vivo*. To identify the conformation of chromatin genome wide, ligation products were subjected to deep-sequencing (Hi-C) (Dixon et al., 2012). Through Hi-C 'topological domains' were identified in the interphase human genome with a median size of 880 Kb. These domains are largely invariant between cell types and even between human and mouse, and may represent a general structural organisation of genomic regions. Together, this data supports a genome organisation made up of large structural loops containing small regulatory loops, which organise the genome at scales of tens to hundreds of kilobases.

The above cytological, differential centrifugation and high-throughput chromatin conformation methods each identify domains of tens to hundreds of kilobases within the human genome. Whether each of these observations reflect the same underlying structure or distinct features of the genome remains unproven and is an important question within the field of chromatin biology, but is beyond the scope of this study.

1.1.5 Chromatin organisation in the nucleus

The distribution of large-scale chromatin structures and whole chromosomes within an interphase nucleus is non-randomly organised *in vivo*. Chromosomes form distinct territories where they are more likely to form intra-chromosomal interactions than inter-chromosomal interactions. Chromosome territories were initially identified cytologically through the use of fluorescent probes for whole chromosomes (Boyle et al., 2011) and subsequently confirmed by Hi-C based molecular studies (Dixon et al., 2012; Lieberman-Aiden et al., 2009). In addition to the intra-chromosomal interactions, a number of reproducible inter-chromosomal interactions have been identified between specific loci by molecular (Dixon et al., 2012; Lieberman-Aiden et al., 2009; de Wit et al., 2013) and cytological techniques (Branco and Pombo, 2006; Kalhor et al., 2012; de Wit et al., 2013). For example, in embryonic stem cells the *Nanog* gene forms inter-chromosomal interactions with other pluripotency genes including *Esrrb* and *Zfp281* in a tissue specific manner (De Wit et al., 2013). It has therefore been suggested that chromosome territories are

composed of a condensed ‘core’ surrounded by a de-condensed ‘corona’ (Bickmore and Van Steensel, 2013). In this model the ‘corona’ contains generally active loci that visibly loops to make contacts with distal regions in the ‘corona’ of the same chromosome or with loci in the ‘corona’ of neighbouring chromosomes (Bickmore and Van Steensel, 2013). A recent study indicates that it is the distribution of the contacts in the ‘corona’, rather than changes in the overall folding of chromosomes, that regulates gene expression through the three dimensional structure of the genome (De Wit et al., 2013). Together this data shows that the three dimensional organisation of chromatin domains within and between chromosomes is a critical component of gene regulation *in vivo*.

1.2 DNA supercoiling

1.2.1 Defining DNA supercoiling

DNA supercoiling is a transient and context dependent structural change in which the helix is over- or under- wound. At the simplest level supercoiling is a mathematical description of topological changes in any circular or constrained-linear double helix. The following section will define the terms for DNA supercoiling used throughout this thesis.

1.2.1.1 Linking number, twist and writhe

The topological properties of the DNA double helix can be defined by equation ‘linking number (L) = twist (T) + writhe (W)’, with linking number corresponding to the number of times one strand crosses the other when the DNA is made to lie on a flat plane and twist/writhe describing the distribution of these cross-overs between the number of helical turns in a length of DNA (twist) and the additional coiling of the double-helix (writhe) (Figure 1.5a). When each strand of the double helix remains unbroken the relationship between twist and writhe can change, but the linking number cannot. For example, in the top left panel of Figure 1a the 210bp relaxed plasmid has a linking number of 22 which is distributed entirely in twist, but a change in twist of -2 requires a compensatory change in writhe of +2 (positive

writhe) in the top right panel. As DNA is a dynamic molecule, it can accommodate different states of twist and writhe, with important biological consequence (see Section 1.2.5).

1.2.1.2 Relaxed DNA double helix

In the case of the DNA double helix the chemical structure determines a preferred helical state which corresponds to the lowest energy form of the molecule. The average helical repeat of linear DNA with ends free to rotate, or nicked circular DNA, is 10.5bp per turn. A covalently closed circular DNA molecule with this repeat is called relaxed DNA and for a given length DNA the linking number of DNA is determined by the equation 'relaxed linking number (L_0) = number of base pairs (N)/10.5' (Figure 1.5b - $L_0 = 20$). Therefore, in relaxed DNA '$L = L_0$'.

1.2.1.3 Positive and negative DNA supercoiling

In supercoiled DNA '$L \neq L_0$', introducing energy that distorts the twist and/or writhe from that of the relaxed double helix. A DNA helix that increases linking number compared to a relaxed helix is positively supercoiled ('$L > L_0$') whereas one that decreases linking number is negatively supercoiled ('$L < L_0$'). The negatively supercoiled plasmid in the bottom panel of Figure 1.5a has a linking number value of 18 compared to 20 in the relaxed plasmid, which can be accommodated by a decrease in twist of -2 or the introduction of -2 negative writhe. The reverse is true of the positively supercoiled plasmid in the top panel of Figure 1.5a (L=22), which distorts the helix through an increase in twist (+2) or a positive change in writhe (+2). Therefore, positive and negative supercoiling are defined by a difference in linking number compared to L_0 and are absolute measures of DNA topology.

1.2.1.4 Under-wound and over-wound DNA

The techniques used in this thesis cannot determine the absolute level of DNA supercoiling and instead determine the relative supercoiling across regions of the

human genome. A relative measure of DNA supercoiling is the specific linking difference (σ), which can be summarised in the equation '$\sigma = (L - L_0) / L_0$'. As σ is proportional to the number of base pairs it can be used to directly compare different regions of DNA, with lower σ values being relatively under-wound DNA when compared to higher σ values (which are relatively over-wound). In this thesis the relative supercoiling of different DNA regions are compared using a chemical probe (see Chapter 4), and the L and L_0 remain unknown. However, the same definition of DNA supercoiling can be applied as we are interested in the difference rather than the absolute levels of supercoiling. As an example, in the model in Figure 1.5c there is a clear difference between the over-wound and under-wound region, even though we do not know the linking number properties. Therefore, in this thesis the definition of under- and over- wound DNA, in terms of specific linking difference, can be represented by 'under-wound < σ < over-wound'.

1.2.1.5 Definitions of supercoiling used in this thesis

For complete clarity Figure 1 models the definitions of supercoiling used in this thesis. Positive and negative supercoiling are absolute measurements of linking number compared to a relaxed circular or constrained DNA molecule and are used only when discussing the work of others (e.g. Baxter et al., 2011; Bermúdez et al., 2010; Kouzine et al., 2008). When discussing the experiments of this thesis, DNA supercoiling is referred to by the relative terms under-wound or over-wound, which are analogous to a comparison of specific linking difference.

In addition, the localisation of DNA supercoils is often referred to as 'upstream' or 'downstream' of a particular position on the DNA, such as a gene or transcribing polymerase. 'Upstream' is defined as 5' and 'downstream' is 3' of the position of interest on the DNA.

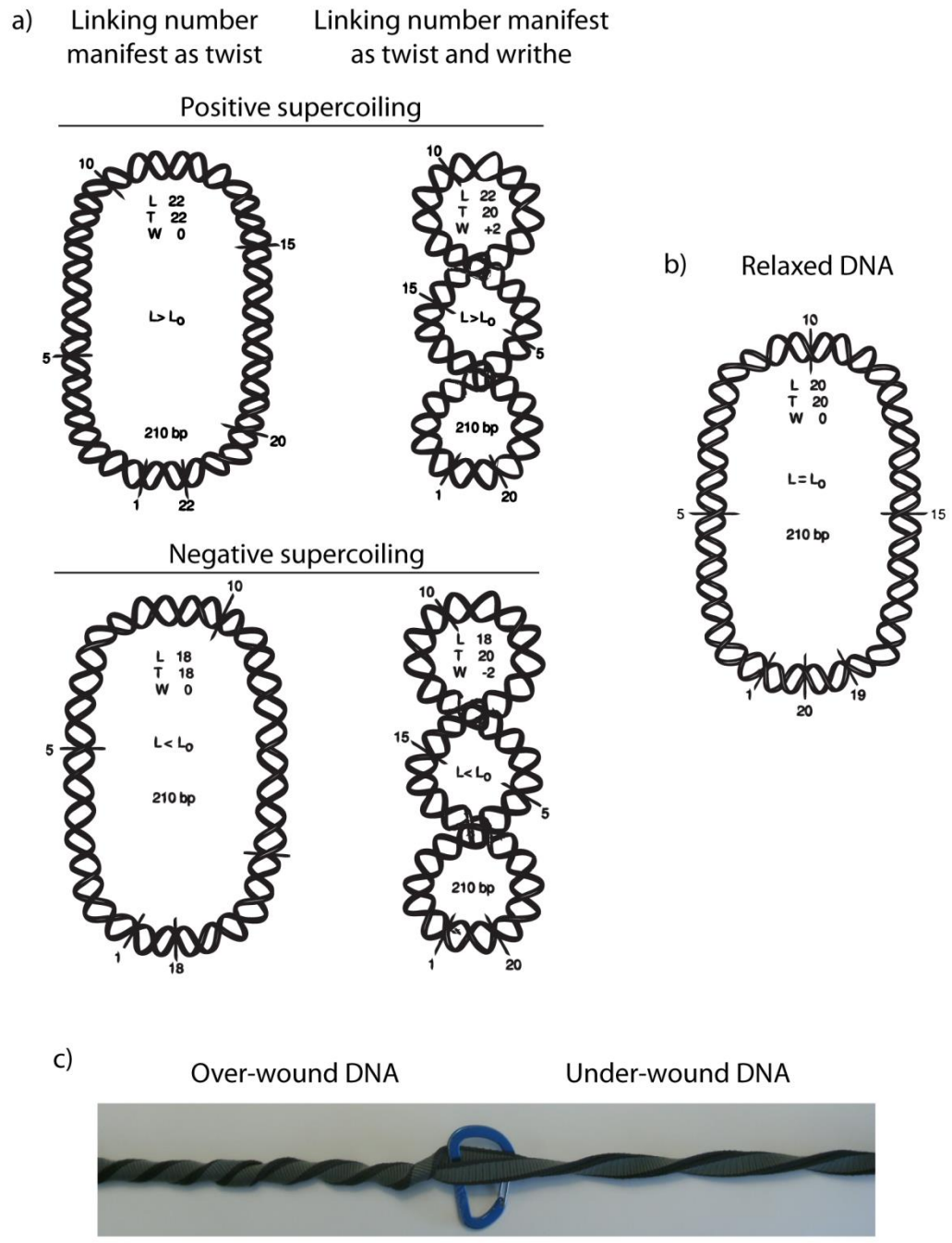


Figure 1.5 Properties of DNA supercoiling. a) Supercoiled and b) relaxed DNA defined in terms of linking number (L), twist (T) and writhe (W). Figure adapted from Sinden (1994). C) DNA supercoiling defined in terms of over-wound and under-wound DNA.

1.2.2 DNA supercoiling in chromatin

The net DNA supercoiling of prokaryotic and eukaryotic genomes is comparable, with one under-wound supercoil per 200 bp in *E. coli* and *Drosophila* (Benyajati and Worcel, 1976). This supercoiling is present in two distinct forms, those restrained in the core particle (or prokaryotic equivalent) and those unrestrained in the linker DNA. The distribution of restrained/unrestrained DNA supercoils is distinct between prokaryotes and eukaryotes, with only ~50% of prokaryotic supercoils being protein bound (Pettijohn and Pfenninger, 1980) compared to almost all eukaryotic supercoils (Bates and Maxwell, 2005; Sinden et al., 1980). This will result in different structural and functional consequences for the packaging and regulation of prokaryotic and eukaryotic genomes.

In prokaryotic genomes the high levels of unrestrained DNA supercoiling and reduced association of DNA with histone-like proteins results in a highly writhed DNA molecule (Postow et al., 2004; Sherratt, 2003). This writhe contributes to the packaging and regulation of the prokaryotic genome. In eukaryotes, our understanding of how unrestrained supercoiling is accommodated into DNA in the context of chromatin is very limited. In part, this stems from a poor understanding of the structures adopted by the higher order chromatin fibre (Section 1.1.3) and the influence of alternative histones and modifications on chromatin fibre plasticity (Gilbert and Allan, 2013). However, it appears that chromatin is highly accommodating of DNA supercoiling (Kouzine et al., 2008; Naughton et al., 2013a) and *in vitro* evidence indicates that the dissipation of supercoils is unhindered by the chromatin macromolecule (Bancaud et al., 2006). Additionally, evidence suggests that yeast chromatin has a greater propensity to absorb DNA supercoiling, whereas the chromatin of higher eukaryotes transmits DNA supercoiling, supporting a model in which chromatin influences the dynamics of DNA supercoiling (Gilbert and Allan, 2013; Morse et al., 1987). Whether unrestrained DNA supercoiling in chromatin is mainly manifest as twist or writhe is currently unknown and remains the object of fierce speculation.

To further understand the distribution of DNA twist and writhe in chromatin it is important to establish the supercoiling density (σ) that is expected to occur within

chromatin associated DNA, and to establish the biophysical limitations of DNA at this supercoil density. To determine the σ value for a 1 kb region of chromatinised DNA placed between highly expressed divergent promoters, Kouzine et al. (2008) stably transfected a plasmid construct into human cells and used the Cre recombinase system to excise DNA circles that trap local DNA supercoils. Using this system the σ value for DNA upstream of two highly transcribed (divergent) promoters was calculated at -0.07. In addition, at this level of σ Kouzine et al. (2008) also show that the FUSE element melts and the FBP and FIR transcription factors bind DNA. Therefore, σ = -0.07 can be generated within the context of human cells by highly expressed divergent promoters, and likely represents an upper estimate to the σ value present upstream of active gene promoters *in vivo*. Subsequent discussion of supercoiling within the context of chromatin will be framed in the context of σ values close to -0.07.

To establish the likely distribution of twist and writhe within chromatin associated DNA, it is important to identify how much twist can be accommodated by naked DNA before a forced transition to writhe or alternative DNA conformations. Through an analysis of electron micrographs of supercoiled plasmids Boles *et al.* (1990) identify that below σ = -0.016 DNA supercoiling is only accounted for by twist and that writhe occurs in a disordered manner. When σ > -0.016 the contribution of twist and writhe to the linking number has a ratio of 1:2, which is consistent at all values tested up to σ = -0.06. Therefore, in naked DNA plasmids twist can account for ~1/3 of under-wound DNA supercoils under biologically relevant levels of DNA supercoiling. To extend these observations and establish the maximum levels of twist a DNA molecule can withstand before forcing a structural transition, Bryant *et al*. (2003) measured twist using a force-measuring optical trap under conditions that precluded the formation of writhe. When the DNA is maintained under tension it can hold a remarkable amount of twist, with σ values of -0.10 and 0.32 observed for under- and over- wound B form DNA respectively. Beyond these levels of DNA supercoiling structural transitions occur, identified through a change in the relative extension of DNA. Therefore, on the balance of experimental data DNA appears capable of constraining significant levels of twist

and writhe, the balance of which is influenced by tension and/or topological constraint within the DNA.

In the context of chromatin it remains unclear whether twist or writhe contributes most to the accommodation of unrestrained DNA supercoils. The wrapping of DNA into nucleosomes limits the localisation of unrestrained supercoils to the linker regions and the formation of a higher order chromatin fibre, with protein-protein interactions between histones in adjacent regions of the fibre, may influence the capacity of the DNA to form writhe. Therefore, a naked DNA plasmid represents a poor model for understanding twist and writhe in the chromatin context. No direct measurement of twist and writhe have been performed on chromatin fibres, but we can infer that the distribution of these properties probably falls somewhere between naked DNA in solution (1:2 twist:writhe) and naked DNA under tension (1:0 twist:writhe). The true distribution of DNA supercoils will be dependent on a number of properties of the chromatin fibre including the level of unrestrained DNA supercoiling, the length of linker DNA, nucleosome stability on supercoiled DNA and differences in chromatin fibre stability.

In support of a significant contribution of twist to the distribution of unrestrained under-wound supercoiling in chromatin, several studies have shown that psoralen binds DNA in eukaryotes in a supercoil dependent manner (Bermúdez et al., 2010; Matsumoto and Hirose, 2004; Naughton et al., 2013a; Saffran et al., 1988; Sinden et al., 1980). The binding of psoralen to under-twisted (as opposed to under-wound, which includes writhe) DNA is thermodynamically favoured, as the intercalation of the drug induces a slight over-twisting of the helix which returns the DNA towards a lower energy ‘relaxed’ state. It is on the basis of this relationship between twist, DNA supercoiling and psoralen binding that the bTMP pull-down developed in our lab is based (Naughton et al., 2013a). However, the presence of twist does not preclude the presence of writhe and it seems likely that both have a significant role in chromatin.

1.2.3 Restrained DNA supercoiling and the linking number paradox

Extensive experimental data has shown that each nucleosome restrains a single under-wound supercoil in the toroidal writhe of the bound DNA helix (e.g. Luger et al., 1997; Richmond and Davey, 2003). However, the length of DNA bound by a nucleosome wraps around the outer surface ~1.8 times and should therefore constrain ~1.8 under-wound supercoils per nucleosome. The difference between the observed and expected constrained DNA supercoils in a nucleosome is called the 'linking number paradox' (Bates and Maxwell, 2005). Extensive structural studies have identified peculiarities in the nucleosome bound DNA structure that partially explain the linking number discrepancy. These include straight DNA sections that do not curve around the nucleosome (Richmond and Davey, 2003) and an over-wound helical structure that cancels out some of the plectonemic writhe through twist (Hayes et al., 1991). Regardless of the mechanism, the DNA supercoils constrained by the nucleosome contribute to the compaction of the genome, and may act as a store of potential energy within the chromatin that is released into the DNA with the dissociation of the nucleosome. This difference between observed and expected serves as a reminder of the hidden complexity of DNA and its influence on DNA-protein interactions, chromatin structure and subsequent gene regulation.

1.2.4 Introduction and removal of DNA supercoils

Interactions between protein and DNA generally involve a twisting, bending or wrapping of the double helix, in each case generating DNA supercoils (Bates and Maxwell, 2005). This is especially true in the case of polymerases, which constantly introduce over- and under-wound supercoils into the DNA as they transcribe or replicate. In addition, the products of DNA replication can coil around one another to produce inter-strand catenanes which must be separated to prevent the formation of DNA knots and for the separation of sister chromatids (Postow et al., 2001). The relief of DNA supercoils and catenanes is performed by DNA topoisomerases, through the transient introduction of nicks and double strand breaks into the DNA

backbone. The activity of these topoisomerases is conserved and essential for all cellular life (Forterre and Gadelle, 2009). In addition, some specialised topoisomerases (e.g. DNA gyrase) have evolved in microorganisms that introduce over- or under- wound supercoils into the DNA though a modified decatenation reaction.

1.2.4.1 Polymerases

1.2.4.1.1 RNA Polymerase

In order to read a single strand of DNA, RNA polymerase requires the localised unwinding of the double helix as it passes along a region of the genome. The size of the polymerase complex, with a combined mass of greater than 2 MDa (He et al., 2013), means that the complex does not rotate with the pitch of the helix during transcription. Therefore, as the polymerase passes between the strands of the double helix, it generates an over-wound helix ahead of itself and an under-wound helix behind (Figure 1.6). This is known as the twin supercoil domain model (Liu and Wang, 1987). In this model RNA polymerase introduces one positive and one negative supercoil per rotation of the double helix (i.e. ~10.5 bp). Experiments *in vitro* and in transfected plasmids *in vivo* confirmed this models validity in biological systems (Giaever and Wang, 1988; Hirose and Suzuki, 1988). The simultaneous introduction of positive and negative supercoils suggests that when the polymerase is removed the over-wound and under-wound supercoils can diffuse back through the DNA and cancel one another out. However, transcription on a circular plasmid in *Xenopus* oocytes generates an under-wound DNA template (Dunaway and Ostrander, 1993), indicating that over-wound DNA is preferentially relieved during transcription. The preferential release of over-wound supercoils will prevent their build up ahead of RNA polymerase, which can slow and eventually prevent transcription by inhibiting DNA strand separation and stalling RNA polymerase (Ma et al., 2013a). Furthermore, high levels of under-wound supercoiling must also be relieved to prevent the formation of single stranded DNA or alternative DNA structures, which can also stall RNA polymerase (French et al., 2011; Ma et al.,

2013a). Therefore, transcription by RNA polymerase II generates DNA supercoils *in vivo* which require an active mechanism for their release.

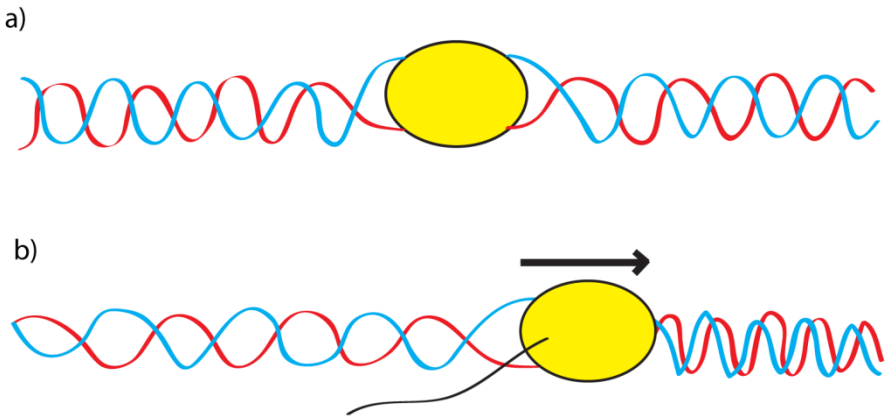


Figure 1.6 Transcription alters DNA supercoils. The binding of the RNA polymerase complex to the DNA requires the localised unwinding of the double helix (a). As transcription proceeds the DNA is over-wound ahead of the RNA polymerase and under-wound behind, generating positive and negative supercoils by the twin supercoil domain model (Liu and Wang, 1987).

1.2.4.1.2 DNA Polymerase

DNA polymerase also requires the localised unwinding of the DNA double helix, generating over-wound supercoils ahead as it replicates the DNA (Postow et al., 2001). However, the newly replicated DNA strands upstream of the polymerase have no net supercoiling, as they are formed by the addition of nucleotides to a single stranded DNA molecule. As supercoiling refers to the linking number of a double helix, a single stranded DNA molecule cannot contain supercoiling and the newly replicated DNA will correspond to the lowest energy form of the double helix. However, if the supercoils generated ahead of the polymerase are not relaxed then the replication fork swivels to unwind the helix ahead of the fork, causing the replicated strands behind the polymerase to coil around one another and form double strand pre-catenanes (Baxter and Aragón, 2010). This formation of precatenanes is most common at the termination of replication where two polymerases, each generating over-wound supercoils, converge on an ever shortening region of DNA. In order to resolve replicated genomes and segregate chromosome appropriately into daughter cells, these pre-catenanes must be removed by the action of type II topoisomerases (see Section 1.2.4.2.2).

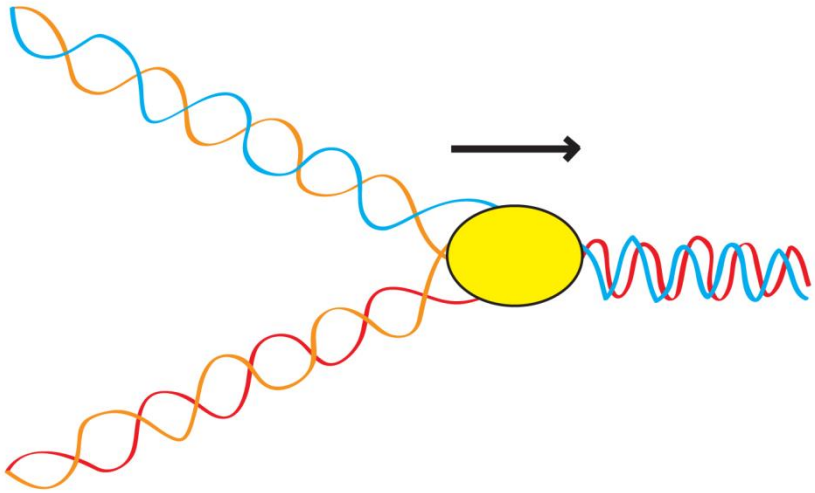


Figure 1.7 DNA replication introduces DNA supercoils. Following a similar mechanism to that of transcription, the processing of DNA polymerase introduces positive supercoils ahead of the replication fork.

1.2.4.2 Topoisomerases

Topoisomerases relieve supercoils and catenanes from the DNA through the introduction of single strand nicks or double strand breaks, classified as type I and type II topoisomerases respectively. Although the precise compliment of topoisomerases varies from species to species, their function is conserved, with all cellular organisms having at least one type I and type II topoisomerase (Forterre and Gadelle, 2009). The topoisomerase classification can be further divided, based on the mechanism of DNA strand cleavage and supercoil relaxation, into topoisomerase IA, IB, IC, IIA and IIB. Confusingly, the 'types' and 'names' of topoisomerase often appear similar, but signify different enzymes. For example, human topoisomerase III enzymes are classified as type IA topoisomerases and topoisomerase IIβ is classified as a type IIA topoisomerase (see Table 1.1). In addition, specialised prokaryotic topoisomerases such as DNA gyrase use ATP to introduce DNA supercoils into the DNA. Identifying the presence, function and distribution of different classes of topoisomerase *in vivo* is necessary to understand how organisms maintain DNA structure, which may itself influence genetic regulation.

Topoisomerase class	Organism and name
IA	Bacterial topoisomerase I and III Yeast topoisomerase III *Drosophila* topoisomerase IIIα and IIIβ Mammalian topoisomerase IIIα and IIIβ
IB	Eukaryotic topoisomerase I Pox virus topoisomerase
IC	*Methanopyus* topoisomerase V
IIA	Bacterial DNA gyrase and topoisomerase IV Phage T4 DNA topoisomerase Yeast DNA topoisomerase II *Drosophila* DNA topoisomerase II Mammailan topoisomerase IIα and IIβ
IIB	Bacterial, plant and archeal topoisomerase VI

Table 1.1 Topoisomerase classification. Topoisomerases 'class' separates proteins based on structure and relatedness. Topoisomerase 'name' is different, representing the name used in the literature for each topoisomerase enzyme.

1.2.4.2.1 Topoisomerase type I

Topoisomerase type IA

Type IA topoisomerases cleave a single strand of the double helix to form a 5'-phosphotyrosyl linkage (Liu and Wang, 1979) creating an enzyme-bridged single stranded gap in the double helix with sufficient width to allow the passing of a second DNA segment (mechanism reviewed in Schoeffler and Berger, 2008). Prokaryotes and archea have topoisomerase I and III enzymes that use the topoisomerase IA mechanism (Srivenugopal et al., 1984), whilst human cells only have topoisomerase III enzymes. The prokaryotic and archeal topoisomerase I enzymes only relax negative supercoils, because the topoisomerase IA mechanism requires a single stranded DNA substrate formed by the localised unwinding of the DNA double helix, a process resisted by positively supercoiled DNA (Kirkegaard and Wang, 1985). The relaxation of negative supercoils by prokaryotic topoisomerase I proceeds to an equilibrium point at which the DNA is still under-wound. This is important for organisms such as *E. coli* where the under-wound structure of the genome is vital for genome packaging and regulation (Postow et al., 2004; Sinden and Pettijohn, 1981). The decatenation activity of prokaryotic topoisomerase I is poor, whereas topoisomerase III is an effective decatenase and a poor relaxase (Hiasa et al., 1994). The difference in activity between topoisomerase I and III is mediated through a 17 amino acid insertion required for the DNA linking activity (Li et al., 2000). In humans topoisomerase IIIα is a single stranded DNA decatenase (Yang et al., 2010) that is important for recombination through the dissolution of alternative DNA conformations (Wu and Hickson, 2003). A further role for topoisomerase IIIα has been identified in the resolution of ultra-fine anaphase bridges between sister chromatids, working together with the DNA helicases BLM and RMI1 in the final stages of chromosome separation (Chan et al., 2007; Yang et al., 2010). The mechanisms and functions of topoisomerase IIIβ are less well studied, although the high sequence similarity in the catalytic core suggests a similar role *in vivo* (Champoux, 2001). Together this data indicates that type IA

topoisomerases in human cells are specialised replication enzymes that work in conjunction with DNA helicases.

One further type IA topoisomerase adapts the 'enzyme-bridged single strand DNA gap' mechanism to introduce positive supercoils into the DNA of hyperthermophilic bacteria and archea. These reverse gyrase proteins are composed of the core catalytic fold of type IA topoisomerases and a helicase-like domain which work together to relax negative supercoils, introduce positive supercoils and preventing the temperature-related degradation of nicked DNA (Schoeffler and Berger, 2008). The positive supercoiling of hyperthermophilic genomes, facilitated by reverse gyrase, is important for genome stability and may help re-nature DNA melted by high temperature conditions (Kampmann and Stock, 2004; Schoeffler and Berger, 2008). This highly specialised protein is an extreme example of the importance of DNA supercoiling in genome stability and regulation.

Topoisomerase type IB

Topoisomerase IB enzymes form a single strand break in which the 3' phosphate forms a phosphotyrosyl linkage with the enzyme (Figure 1.8a), whilst the DNA downstream of the break rotates under the free energy of the unrestrained supercoiling in the helix (Figure 1.8b)(Champoux and Dulbecco, 1972; Koster et al., 2005). Type IB enzymes include topoisomerase I in eukaryotes, poxvirus and some bacteria. The topoisomerase protein clamps tightly around the DNA (Champoux, 2001; Leppard and Champoux, 2005), forming several protein-DNA interactions which slow the rotation and position the break for re-ligation (Koster et al., 2005; Stewart et al., 1998). This mechanism releases a single supercoil from the helix per revolution, with multiple revolutions occurring between strand break and religation in a torsion-force dependent manner (Koster et al., 2005). The controlled rotation mechanism removes both over-wound and under-wound supercoils from the DNA, which are necessary for topoisomerase binding (Madden et al., 1995). However, the relaxation of positive supercoils by topoisomerase I occurs around 50 times faster than negative supercoils (Frøhlich et al., 2007) and it has been suggested that positive supercoils may specifically recruit topoisomerase I (Leppard and Champoux, 2005).

Topoisomerase I is enriched in transcribed genes in *Drosophila* (Gilmour et al., 1986), co-localises with RNA polymerase II associated transcription factors (Kretzschmar et al., 1993) and enzyme activity (Stewart et al., 1990), is associated with RNA polymerase I and rDNA transcription (Christensen et al., 2004), localises to replication forks (Leppard and Champoux, 2005) and is linked to open chromatin and gene expression (Durand-Dubief et al., 2010; Filion et al., 2010) (Section 3.1). However, our understanding of the relevance of topoisomerase I co-localisation and function is currently restricted as the precise distribution has not been mapped for any significant portion of the human genome. Unsurprisingly for a protein intimately linked with transcription and replication, topoisomerase I is required for viability in fly and mouse (Lee et al., 1993; Morham et al., 1996), but yeast topoisomerase I knock-outs are viable (Kim and Wang, 1989). This may be a reflection of differences in gene size, developmental complexity or the complement and function of topoisomerase enzymes in different organisms.

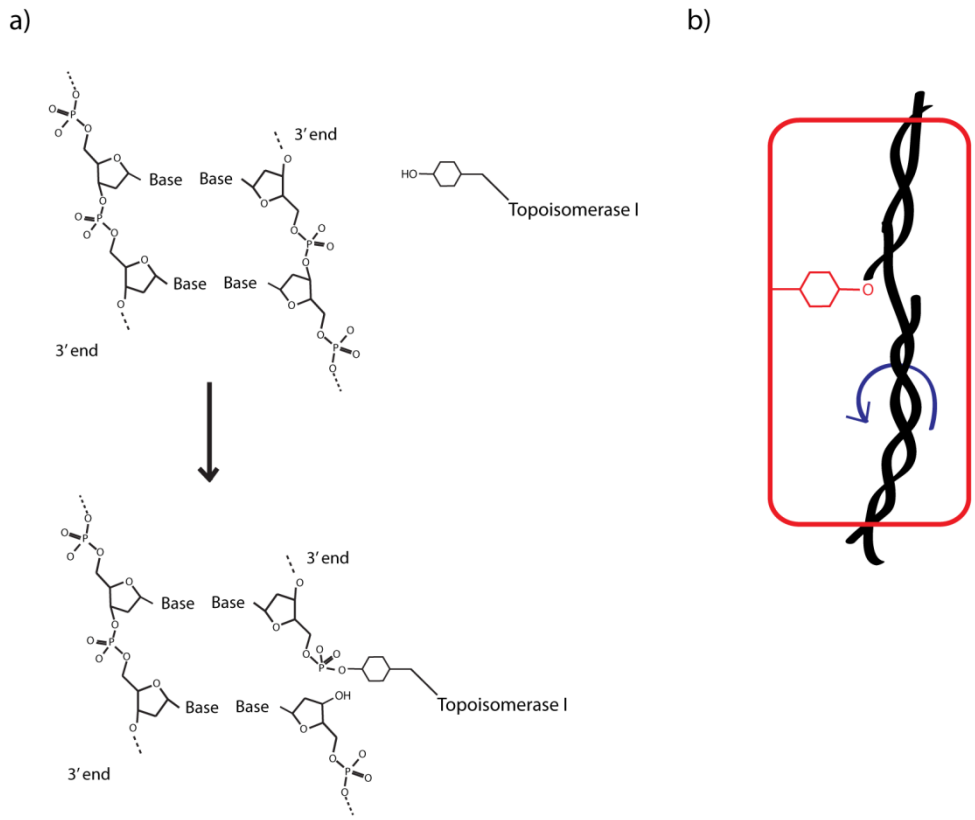


Figure 1.8 Topoisomerase IB mechanism. a) The formation of a 3' phosphotyrosyl linkage between the topoisomerase enzyme and DNA stabilises the enzyme induced nick. b) Schematic of the controlled rotation mechanism within the topoisomerase enzyme.

Topoisomerase type IC

Topoisomerase V is a type I topoisomerase that is structurally and phylogenetically distinct from other identified topoisomerases, but is mechanistically similar to topoisomerase IB enzymes (Schoeffler and Berger, 2008). This topoisomerase has only been identified in thermophylic archea of the genus *Methanopyus* that inhabit hydrothermal vents (Forterre and Gadelle, 2009) and is most notable for performing its enzymatic activity optimally at 108°C (Kozyavkin et al., 1995). The convergent evolution of topoisomerase IB and IC enzymes further supports the importance of DNA supercoil relief through transient single strand breaks in cellular organisms.

1.2.4.2.2 Topoisomerase type II

Topoisomerase type IIA

Topoisomerase II enzymes are complex multi-subunit molecular machines which form enzyme-bridged breaks in double stranded DNA (Figure 1.9a/b), through which a second double stranded DNA helix is passed (Figure 1.9b) (Liu et al., 1980). In general, this releases positive and negative DNA supercoils two at a time and can decatenate inter-wound helices and DNA knots. Type IIA enzymes include eukaryotic topoisomerase IIα and IIβ, viral and bacteriophage topoisomerase II, bacterial and archeal DNA gyrase and bacterial topoisomerase IV (Champoux, 2001; Schoeffler and Berger, 2008). The functions of topoisomerase IIA enzymes are wide ranging, with forms that preferentially relieve supercoils, introduce supercoils and decatenate.

Topoisomerase II enzymes are type IIA topoisomerases that are present in eukaryotes, viruses and bacteriophages (Table 1.1). They function both to relieve supercoils and decatenate DNA strands, with many organisms expressing different forms specialised to these different functions. One study in yeast supports topoisomerase II as the major relaxase in chromatin, as opposed to naked DNA

(Salceda et al., 2006). Using a yeast minichromosome system they show that topoisomerase II is 5 times more efficient than topoisomerase I at relieving supercoils from a chromatinised template. This observation seems at odds with the published distributions of these enzymes in higher eukaryotes, with topoisomerase I associated with actively transcribed genes and topoisomerase II associated with AT-rich regions of the genome (Gilmour et al., 1986; Käs and Laemmli, 1992). The authors account for this apparent contradiction by postulating that chromatin may buffer the driving torque required for topoisomerase I, therefore topoisomerase II is generally more efficient on a chromatin template, except in the immediate vicinity of a polymerase where molecular crowding prevents the formation of writhe and generates high levels of twist for release by topoisomerase I. Whether this model is specific for yeast, or even just yeast minichromosomes, has not been established. Therefore, on the basis of this model topoisomerase II may be more important in the vicinity of transcribed regions for the release of DNA supercoils, and should be considered in an interpretation of topoisomerase II distribution *in vivo*.

In higher eukaryotes topoisomerase IIα and topoisomerase IIβ are evolutionarily conserved proteins that are likely to perform different functions within the cell (Champoux, 2001). Topoisomerase IIα is essential for chromosome segregation and performs a redundant role in chromosome condensation in human cells (Carpenter and Porter, 2004), indicating a role in the decatenation of DNA. Supporting this, the expression of topoisomerase IIα varies through the cell cycle, with expression highest in S phase through to G2-M phase where decatenation is critical (Woessner et al., 1991). The observation that topoisomerase IIα preferentially binds positively supercoiled DNA (McClendon et al., 2005) may indicate a further role in DNA replication, resolving the positive supercoils ahead of the polymerase and the catenanes behind in a cell cycle dependent manner. Topoisomerase IIβ, on the other hand, is expressed throughout the cell cycle (Woessner et al., 1991) and shows no preference for binding supercoiled DNA (McClendon et al., 2005). If topoisomerase II enzymes are the more efficient relaxase in chromatin, then it is likely that topoisomerase IIβ performs this role throughout the cell cycle in higher eukaryotes. One function of topoisomerase IIβ is in the regulation of transcription initiation and repression (Ju et al., 2006; Lyu et al., 2006; McNamara et al., 2008). However, cell

lines in which topoisomerase IIβ has been knocked out are viable (Nitiss, 2009a). This indicates that unlike topoisomerase IIα, topoisomerase IIβ is not required in the cell cycle or for survival. The basis of the functional difference between topoisomerase IIα and IIβ has been of much interest, particularly as they share a 72% amino acid sequence identity, and must lie in the more variable non-catalytic C-terminal domain (Champoux, 2001). Whether this affects the spatial distribution of topoisomerase IIα and IIβ in the genome is unknown. Cytological studies have indicated that in human chromosomes topoisomerase II enzymes form one component of the chromosome scaffold in metaphase (Earnshaw and Heck, 1985) and form the base of large chromatin loops (Paulson and Laemmli, 1977). Topoisomerase II may form a structural component of these loops and regulate DNA supercoiling within. However, in these studies the majority of protein was stripped from the metaphase chromosomes, therefore the scaffold may be just one example of a feature associated with topoisomerase II. For example, it has been shown that topoisomerase IIβ is enriched at some promoters and it has been suggested that this distribution may be a general feature of promoters in the genome (Kouzine et al., 2013; Lyu et al., 2006; Sano et al., 2008). In prokaryotes type IIA topoisomerases either relieve DNA supercoils and catenanes or introduce negative supercoils in an ATP dependent manner. Prokaryotic topoisomerase IV has a similar structure and function to eukaryotic topoisomerase II enzymes, but with a preference for positively supercoiled and catenated substrates (Crisona et al., 2000; Zechiedrich et al., 1997). DNA gyrase, on the other hand, adapts the type II topoisomerase mechanism to actively introduce negative supercoils into the DNA (Gellert et al., 1976). The mechanism of DNA gyrase activity involves wrapping one strand of DNA around the enzyme so that the same strand is both broken and passed through itself, introducing two negative supercoils per reaction (Schoeffler and Berger, 2008). DNA gyrase activity is required for DNA replication in many prokaryotes, but is absent from human cells, and is therefore a major target for antibiotics (Engle et al., 1982). Together, the activity of topoisomerase IV and DNA gyrase maintain the high levels of negative supercoiling present in the *E. coli* genome. This is a further example of topoisomerases regulating general genome structure and function through DNA supercoiling.

Topoisomerase type IIB

The single type IIB topoisomerase identified thus far is topoisomerase VI, which is a distant relative of type IIA topoisomerases present in archea, plants and some bacteria (Forterre and Gadelle, 2009; Schoeffler and Berger, 2008). Topoisomerase VI has a similar mechanism to type IIA topoisomerases, relaxing positive and negative supercoils and decatenating through a strand passage mechanism. The discovery of topoisomerase VI is relatively recent for a type II topoisomerase (Bergerat et al., 1994) and little is known of its distribution and function.

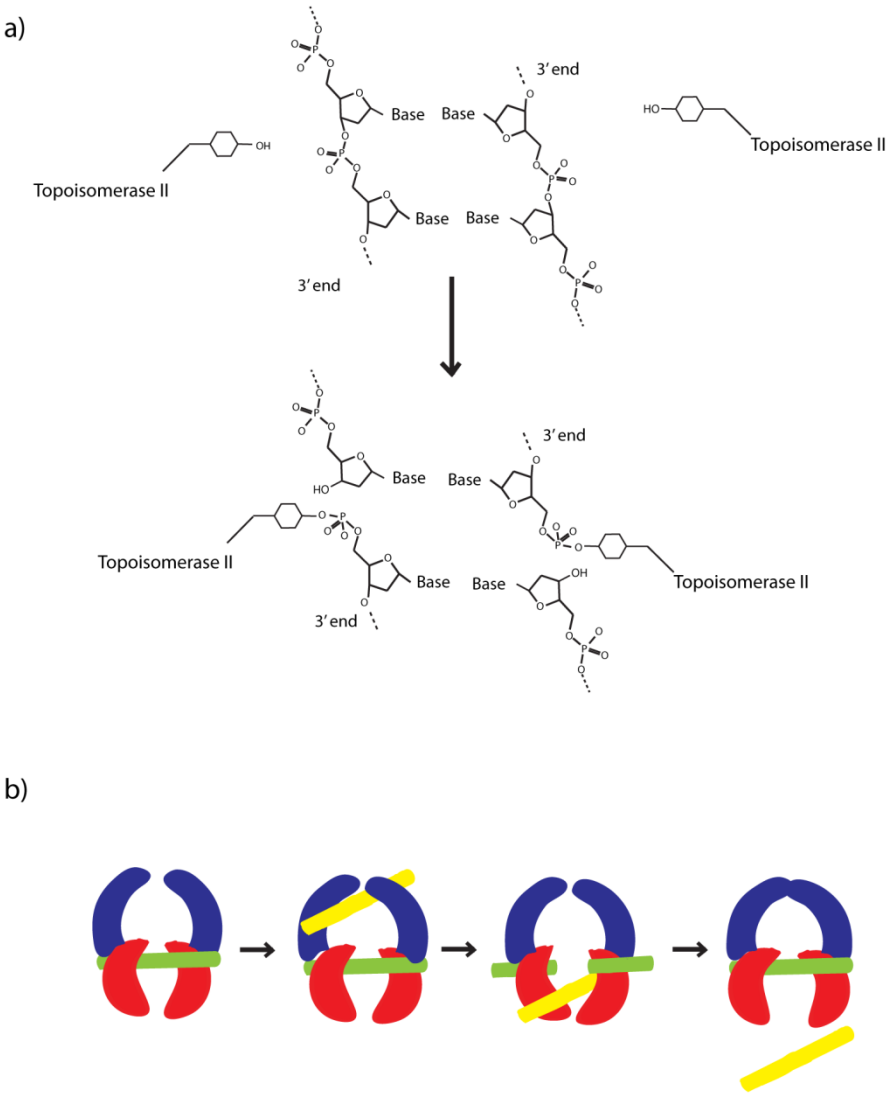


Figure 1.9 Topoisomerase IIA Mechanism. a) The formation of 3' phosphotyrosyl linkages between the topoisomerase enzyme and both strands of a DNA double helix. b) Schematic of the strand passage mechanism. The green DNA strand is broken through the mechanism in a) and a second double stranded DNA is passed through the gap.

1.2.5 DNA supercoiling *in vivo*

In the cellular context it is conceivable that DNA supercoiling introduced by transcription and replication is completely removed from the DNA by topoisomerase enzymes to leave the genome in a relaxed state. In prokaryotes this is clearly not the case, with genomes maintaining significant unconstrained negative supercoils (Sinden and Pettijohn, 1981; Sinden et al., 1980). More recently domains of unrestrained supercoils have been observed in the DNA of yeast, fly and human (Bermúdez et al., 2010; Matsumoto and Hirose, 2004; Naughton et al., 2013a). This indicates that DNA supercoiling may have a biological role in eukaryotes similar to that observed in prokaryotes.

1.2.5.1 Biological roles of DNA supercoiling

Supercoiling introduces energy into the DNA which can be dissipated through the fibre or localised to distort the structure of the double helix. Both of these manifestations of DNA supercoiling have been shown to influence transcription by a number of mechanisms, including the selective binding of transcription factors to altered DNA structures (e.g. Kouzine et al., 2008), an increased transcription efficiency on an under-wound DNA template (Dunaway and Ostrander, 1993; Weintraub et al., 1986) and the large-scale decompaction of chromatin structure (Matsumoto and Hirose, 2004; Naughton et al., 2013a). The dissipation of DNA supercoiling through the DNA fibre has been shown to generate domain scale (Matsumoto and Hirose, 2004; Naughton et al., 2013a) or genome-wide (Sinden and Pettijohn, 1981; Sinden et al., 1980) under-wound DNA structures in eukaryotes and prokaryotes respectively. Early studies of DNA supercoiling in transfected plasmids identified that under-wound DNA is transcribed more efficiently than linear or nicked DNA in *E. coli* (Weintraub et al., 1986) or *Xenopus* oocytes (Dunaway and Ostrander, 1993), providing the first evidence of the regulatory potential of DNA supercoiling *in vivo*. Subsequent genome wide analyses in prokaryotes identified that many, but not all, genes are regulated by DNA supercoiling (Lyu et al., 2006; Peter et al., 2004). Taken together, these studies support the direct regulation of genes by DNA supercoiling in plasmid and prokaryotic systems. In eukaryotes there

is a strong correlation between under-wound DNA and gene expression, supporting a similar regulatory mechanism. For example, domains of under-wound DNA on human chromosome 11 are significantly enriched for transcription, RNA polymerase II, open chromatin and DNaseI hypersensitivity when compared with relaxed or over-wound domains (Naughton et al., 2013a). The inhibition of transcription by RNA polymerase II leads to a significant re-organisation of DNA supercoil structure at transcriptionally active regions, which corresponds to a more compact large-scale chromatin structure. Furthermore, expression of *Drosophila hsp70* following heat shock coincides with a localised under-winding of DNA structure and decondensation of the chromatin structure (Matsumoto and Hirose, 2004). Together, these studies support a general mechanism in which domain scale DNA supercoil structure, transcription and chromatin accessibility are all inextricably linked for large-domains across the human and fly genome, a mechanism first proposed at the chicken *β-globin* locus by Villeponteau et al. (1984). One particularly attractive model is that these supercoil domains co-transcriptionally regulate neighbouring genes, with the activation of one gene having a direct influence on others through the dissipation of DNA supercoils. This could account for the 'transcription ripple effect' observed by Ebisuya et al. (2008), in which the activation of a gene leads to a delayed increase in expression of neighbouring genes. This ripple effect occurs over domains of ~100kb, remarkably similar in size to the supercoiling domains we have observed. Also, the co-expression is independent of gene pair orientation, supporting the domain scale DNA superoil distribution (Naughton et al., 2013a), rather than the more focal supercoils generated by divergent transcription (Naughton et al., 2013b). Validating this model is beyond the scope of this thesis, but may provide a mechanism by which the observations of Ebisuya et al. (2008) can be explained.

In addition to domains scale changes in DNA supercoiling, transcription alters DNA supercoiling at promoter regions. Transcription at gene promoters is not restricted to the production of full-length mRNA, with extensive short RNA transcription occurring close to the TSS in the sense and anti-sense direction (Core et al., 2008; Preker et al., 2008; Seila et al., 2009). In mammals ~80% of transcribed genes exhibit divergent transcription at the promoter, generating under-wound DNA supercoils into the promoter region. This transcription could represent 'sloppy'

transcription initiation events or the generation of an under-wound DNA structure at the promoter region could be functional, as suggested by Seila et al. (2009). Supporting a model whereby the transcription of short RNAs alters promoter DNA structure, our lab has identified changes in promoter DNA supercoiling that are dependent on the transcription of short RNAs but not long RNAs (Naughton et al., 2013b). The resulting under-wound DNA structure may facilitate promoter unwinding and transcription.

The identification of domains of DNA supercoiling in prokaryotes and eukaryotes indicates a conserved regulatory mechanism. To establish how under-wound DNA facilitates transcription, transcription rate was measured for linear and supercoiled circular DNA containing the *Bombyx mori* fibroin gene at three different steps: transcription initiation, conversion to an elongation complex and subsequent elongation (Tabuchi and Hirose, 1988). In this system, transcription initiation is much more rapid on an under-wound template compared to a linear DNA, whereas the rate of elongation is unaffected. This is consistent with observations that under-wound DNA facilitates the binding of transcription factors and the pre-initiation complex at many gene promoters (Mizutani et al., 1991a, 1991b; Tabuchi and Hirose, 1988). This suggests that large-DNA supercoil domains may be set up and maintained for the preservation of DNA structure at gene promoter regions. In support of this, promoter regions have been shown to have a distinct, generally under-wound structure *in vivo* (Jupe et al., 1993; Kouzine et al., 2013; Ljungman and Hanawalt, 1992, 1995; Naughton et al., 2013a). Furthermore, the generation of an under-wound DNA structure by bivalent transcription at active gene promoters can additionally regulate promoter supercoiling at mammalian genes (Naughton et al., 2013b; Seila et al., 2009).

There are several possible mechanisms through which under-wound DNA facilitates transcription initiation and gene expression, although a general model for DNA supercoil dependent regulation is yet to emerge. For example, under-wound DNA could promote the localised unwinding of the double helix upstream of the transcription start site, which is necessary for the production of a transcription bubble and subsequent transcription. Conversely, the localised over-winding of DNA would

make unwinding at the transcription start site less energetically favourable and may negatively regulate gene expression. Other supercoil dependent transitions in DNA structure at promoter regions include the melting of the Far UpStream Element (FUSE) of the human *c-myc* gene (Kouzine et al., 2004, 2008). The FUSE element is located 1.2 kb upstream of the transcription start site and unwinds when the gene is transcribed. This occurs on both a circular and linear template, indicating that DNA supercoils can build up in the absence of local constraint. The melting of the FUSE element by DNA supercoiling provides a substrate for the FUSE-binding protein (FBP) and the FUSE-interacting repressor (FIR) which modulate *c-myc* expression. The structure of FUSE is highly dynamic, with the unwound DNA structure lost within 10 seconds following transcription inhibition. This provides the first specific example of DNA supercoiling as a real-time sensor and regulator of transcription and it is likely that many similar examples will be identified in the future. Other general examples of DNA structures that occur preferentially on an under-wound template, and may therefore act as sensors/regulators of gene expression, include alternative DNA structures and Z-form DNA. Alternative DNA structures such as cruciforms and G-quadruplexes require the local unwinding of the DNA double helix to allow non-'Watson and Crick' base-pairing. On the other hand, Z-form DNA is stabilised by under-wound supercoils without a loss of 'Watson and Crick' base interactions. The formation of these supercoil dependent DNA structures is most common in the vicinity of promoter elements, with G-quadruplex and Z-DNA motifs concentrated in these regions (Cer et al., 2011). Furthermore, the identification of proteins that specifically bind these structures, such as ADAR1 with its specific Z-DNA binding domain(Herbert et al., 1997), supports the hypothesis that these DNA structures have regulatory potential. Therefore, under-wound DNA supercoiling promotes the unwinding of the DNA double helix, particularaly at specialised motifs, which can influence expression directly, or through the formation of alternative DNA structure, or through the recruitment of regulatory proteins that bind altered DNA structures.

1.2.5.2 Distribution of DNA supercoiling *in vivo*

1.2.5.2.1 Domain scale distribution of DNA supercoiling

As previously mentioned, the functional distribution of DNA supercoiling was first identified in *E. coli*, where the genome is maintained in a state of net unconstrained negative supercoils (Sinden et al., 1980). The distribution of negative supercoils in the *E. coli* genome is dependent on the underlying looped rosette structure (Neidhardt and Curtiss, 1999), which organises the genome into around 50 distinct supercoiling domains of around 100 kb (Sinden and Pettijohn, 1981). Variable supercoiling within each looped domain is thought to contribute to the coordinated expression of neighbouring genes. To test this in *E. coli* Peter et al. (2004) identified 300 'supercoiling sensitive genes' by analysing gene expression changes following topoisomerase inhibition. Monitoring the expression of these genes following DNA relaxation by restriction enzyme digestion of DNA *in vivo* identified that genes within ~11 kb of each other are typically co-transcriptionally regulated in a supercoiling dependent manner (Postow et al., 2004). This range can encompass several genes, which in *E. coli* have an average length of ~1 kb and are organised at a high gene density (Blattner et al., 1997). These 11 kb domains are substantially smaller than the 50-100 kb topological loops, supporting the model proposed by Jeong et al. (2004) that the *E. coli* genome is structured as large scale supercoiling domains (~100 kb) and more transient smaller scale supercoiling domains (~10 kb) that co-exist to regulate transcription at different levels. The transience of the smaller loops is further supported by the absence of defined boundaries in the 10 kb domains (Postow et al., 2004). Together this data indicates that in *E. coli*, domains of altered DNA supercoiling exist at different scales and regulate gene expression *in vivo*.

The functional distribution of supercoiling in eukaryotes has been more difficult to establish, due to the added complexity of different levels of chromatin organisation. Early studies comparing the supercoiling in *Drosophila* and *E. coli* genomes concluded that no net unrestrained supercoiling is present in eukaryotic genomes (Sinden et al., 1980). Instead, the DNA supercoiling of eukaryotic genomes was thought to be predominantly constrained by histone proteins. Stripping histone

proteins from *Drosophila* DNA leaves one negative supercoil per 200 bp, the same net supercoil density as in *E. coli*, indicating a conserved net supercoil density between species (Benyajati and Worcel, 1976). Furthermore, the partial relaxation of constrained supercoils by DNAaseI in histone depleted *Drosophila* genomes identifies that distinct ~85 kb domains of DNA occur in eukaryotes (Benyajati and Worcel, 1976), similar to the large scale structural domains observed in *E. coli* (Sinden and Pettijohn, 1981). This data suggests a conserved large-scale genome structure, but that unconstrained DNA supercoiling and the associated regulatory mechanisms are a primarily prokaryotic phenomenon.

The identification of micro-domains of negatively supercoiled DNA at eukaryotic gene promoter and enhancer elements using psoralen suggested that the importance of unrestrained DNA supercoils in eukaryotic gene regulation had been underestimated by previous studies (Jupe et al., 1993; Ljungman and Hanawalt, 1992, 1995). Psoralen is a planar molecule that preferentially intercalates with under-wound DNA (Bermúdez et al., 2010) and forms covalent cross-links with the DNA upon UV exposure (Cech and Pardue, 1977) (Section 5.1). Localised enrichments for under-wound DNA were identified at promoters in human (Ljungman and Hanawalt, 1992), fly (Jupe et al., 1993) and hamster (Ljungman and Hanawalt, 1995). Extending the use of psoralen to identify the relative distribution of unrestrained DNA supercoils across eukaryotic genomes identified large scale domains similar to those observed in prokaryotes. In general, the preferential intercalation of psoralen is two-fold enriched in under-wound DNA (Kouzine et al., 2013; Naughton et al., 2013a) and therefore remained undetectable at most loci prior to immunofluorescence studies or high resolution, large scale immunoprecipitation/array based approaches. Initial observations of psoralen enrichment by immunofluorescence in *Drosophila* polytene chromosomes identified ~150 domains of enriched psoralen binding that are lost upon nicking or transcription inhibition (Matsumoto and Hirose, 2004). In this study the authors indicate that the under-wound DNA accounts for only a minority of the genome, although their assay is based on the lowest observable threshold of fluorescence and is likely to identify only the most intense peaks of under-wound DNA. To identify the relative distribution of supercoiling at high resolution, the relative enrichment of psoralen

was measured across the genome of yeast using tiling microarrays (Bermúdez et al., 2010). Yeast cells were incubated with psoralen, UV cross-linked and the extracted DNA denatured and digested with exonucleases to leave only the DNA cross-linked by psoralen molecules, which were hybridised to tiling microarrays. Using this technique a domain scale change of ~100 kb was observed between wild type and topoisomerase depleted cells, indicating that topoisomerases can remodel the DNA structure of large domains in eukaryotes. This analysis supports the presence and regulation of unrestrained DNA supercoils in eukaryotic cells, but is at a relatively low resolution, with one probe every 2 kb (i.e. less than one per gene), and fails to take into account the well documented sequence preference of psoralen (e.g. Kanne et al., 1982). To further investigate the distribution of DNA supercoiling in the human genome our lab has developed a different approach based on the pull-down of a bTMP molecule (Naughton et al., 2013a). In this methodology bTMP (Saffran et al., 1988) is incubated with live cells, photo-crosslinked with UV light and the DNA isolated, purified, pulled-down with avidin and hybridised to Agilent tiling microarrays. The data is truly 'high resolution', with 50 bp probes spaced every 100 bases, and covers 20 Mb of the human genome encompassing gene rich, gene poor and other regions hypothesised to be influenced by DNA supercoiling state. Domains of over-wound, under-wound and stable DNA were identified across the loci, with a median size of ~100 kb. These DNA supercoiling domains are relieved with the introduction of DNA nicks by bleomycin and are maintained by transcription and topoisomerase activity. Under-wound DNA domains correspond to transcriptionally active open chromatin whilst over-wound domains are transcriptionally inactive, supporting the *in vitro* models of transcription regulation. Together, this data in fly, yeast and human indicates that unrestrained supercoils are pervasive in eukaryotic genomes and are maintained by transcription and topoisomerase activity.

1.2.5.2.2 Organisation of DNA supercoiling around gene promoters

The presence of unrestrained supercoiling in the genomes of eukaryotes has reignited an interest in the role of DNA supercoiling in gene regulation. The regulation of transcription initiation at gene promoters has been demonstrated to be the primary influence of DNA supercoiling (Section 1.2.5.1) supported by several studies that have identified an under-wound DNA structure at active and poised promoters (Jupe et al., 1993; Ljungman and Hanawalt, 1992, 1995). To identify the general DNA structure of gene promoters, two recent studies have performed a meta-analysis of psoralen distribution over a large number of promoter regions (Kouzine et al., 2013; Naughton et al., 2013a). Our lab performed a meta-analysis of the promoters of 584 human genes, identifying a general under-wound DNA structure which is more enriched at transcriptionally active genes (Naughton et al., 2013a). This under-wound structure is lost when the DNA is nicked with bleomycin and in the presence of transcription or topoisomerase inhibitors. Supporting this, a recent analysis of 445 gene promoters also identifies an enrichment for under-wound DNA at gene promoters that is more pronounced in transcriptionally active genes (Kouzine et al., 2013). Together these studies support a specialised DNA supercoiling structure at gene promoters, but so far the scope is too limited to uncover specific classes of promoter structure *in vivo*.

1.3 DNA structure, topoisomerases and disease

1.3.1 DNA structure and disease

DNA supercoiling and topoisomerases are important factors in human disease. The best defined example of the relationship between DNA structure and disease is in neurological disorders characterised by tri- and tetra-nucleotide repeat expansions, including myotonic dystrophy type 1 (DM1), myotonic dystrophy type 2 (DM2), fragile X-syndrome and Friedreich ataxia (Bacolla and Wells, 2009). In these diseases a massive expansion of repeat sequences occurs at rare fragile sites, which do not encode genes but have the potential to form hairpin, triplex and quadruplex structures. The instability and mutagenic effects associated with this repeat expansions have been shown to be a direct consequence of alternative DNA structure formation (Wojciechowska et al., 2006). Furthermore, DNA supercoiling can build up in the inherently flexible triplet repeat regions and it has been suggested that DNA supercoiling mediates the expansion of these repeat regions (Gellibolian et al., 1997).

Other unstable sequences that are associated with alternative DNA structure include common fragile sites (CFSs). CFSs sites are chromosome regions that reproducibly form breaks, constrictions, gaps or fail to compact upon partial replication inhibition (Lukusa and Fryns, 2008). DNA within fragile sites is highly flexible and has a propensity to form alternative DNA structures (Burrow et al., 2010). The high DNA flexibility at CFSs may allow the build-up of DNA supercoils, as hypothesised at rare fragile sites (Gellibolian et al., 1997), which could abrogate topoisomerase function causing strand breakage (Lukusa and Fryns, 2008). CFSs form a normal component of the chromatin structure and are conserved in mouse and chicken (Helmrich et al., 2006; Le Tallec et al., 2013), although why a conserved chromatin structure forms frequent breaks and rearrangements under replication stress is not well understood. In human cells there 88 documented common fragile sites (CFSs), which are associated with deletions and rearrangements in various cancer types, autosomal recessive juvenile Parkinsonism and Duchenne muscular dystrophy (Lukusa and Fryns, 2008; Mitsui et al., 2010). The complement of CFSs expressed following partial replication stress varies between cell types, with FRA3B and

FRA16D being the most common CFSs in lymphoblastoid cells and sites at 3q13.3 and 1p31.1 being most common in fibroblasts (Le Tallec et al., 2011). This indicates that the expression of CFSs has an epigenetic component, with sequence, replication and other molecular characteristics indicating a role for DNA supercoiling (Section 5.1). To better understand how DNA supercoiling influences CFSs, it is therefore important to map the distribution and maintenance of DNA supercoiling at these loci.

1.3.2 Topoisomerases and disease

Topoisomerases have been implicated in scleroderma, an auto-immune disease characterised by lesions of the skin and in severe cases internal organs (Gabrielli et al., 2009). In patients with the Scl-70 disease sub-type, antibodies to topoisomerase I are found at high concentrations in the blood (Guldner et al., 1986). In other patient sub-types, antibodies to other proteins important for DNA supercoiling and chromatin structure have been identified including RNA polymerase, topoisomerase II and centromeric proteins (Gabrielli et al., 2009). Why these patients have antibodies to these proteins is unknown, although the different symptoms and features of genome instability presented by patients with different autoantibodies suggests a direct role for the antibody in the disease phenotype (Gabrielli et al., 2009; Jabs et al., 1993). For example, peripheral blood lymphocytes from scleroderma patients with anti-centromere antibodies show significantly higher aneuploidy than patients without anti-centromere antibodies or controls (Jabs et al., 1993). However, a specific molecular or cytological phenotype for scleroderma patients with anti-topoisomerase antibodies has not been determined and the function of these antibodies in scleroderma patients is unknown.

Topoisomerases have also been critical in the treatment of diseases, in particular cancer. Topoisomerase activity is essential in rapidly dividing cells (Section 1.2.4.2), therefore topoisomerase inhibitors have been used extensively as a target in chemotherapy (Nitiss, 2009b; Pommier, 2006). Topoisomerase poisons act in several ways to prevent relaxase and decatenase activities by inhibiting topoisomerase I or topoisomerase II enzyme activity. The topoisomerase I inhibitors based on camptothecin and indenoisoquinolines trap the enzyme in a covalent

complex with DNA that prevents the religation of the nicked DNA helix, trapping the enzyme on the DNA (Pommier, 2006). The mechanisms of topoisomerase II inhibition are more varied, with molecules that prevent the opening of the protein to allow the binding of the first (aclarubicin) or second (ICRF-187) DNA strand, those that prevent the formation of a double strand break (melbarone) or those that prevent the release of the strand passed through the double strand break (etoposide) (Nitiss, 2009a). In the laboratory, these drugs have been particularly useful for identifying the molecular characteristics of topoisomerase enzymes.

1.4 Thesis aims

The relationship between DNA supercoiling, transcription and gene expression is well characterised *in vitro* and in transfected plasmids (Dunaway and Ostrander, 1993; Ma et al., 2013a; Weintraub et al., 1986), but the distribution and function within the human genome is largely uncharacterised. DNA with a more under-wound structure has a higher rate of transcription and the polymerase pauses less frequently and for a shorter duration (Ma et al., 2013a). Furthermore, under-wound DNA promotes the binding of transcription factors, the formation of a pre-initiation complex and transcription initiation (Kouzine et al., 2008; Mizutani et al., 1991a, 1991b; Tabuchi and Hirose, 1988). These properties indicate that under-wound DNA supercoiling can facilitate transcription initiation and gene expression.

The characterised presence of DNA supercoiling in the human genome was, until recently, limited to a small number of promoter regions (Ljungman and Hanawalt, 1992). This was due to insufficient large-scale techniques for the study of DNA structure *in vivo*. Recent work in our lab has used bTMP pull-down and microarray analysis to identify large DNA supercoil domains and smaller scale promoter enrichments in the human genome (Naughton et al., 2013a). Using this technique it was demonstrated that both domain-scale and promoter-scale DNA supercoil distributions are dependent on transcription and topoisomerase activity. However, to date the distribution of topoisomerase I, IIα and IIβ enzymes in the human genome have been poorly characterised (Cowell et al., 2012; Khobta et al., 2006; Kouzine et al., 2013). I have undertaken a detailed study of topoisomerase distribution by chromatin immunoprecipitation and microarray analysis (ChIP-chip) (Chapter 3). I have analysed the distribution of topoisomerases over large-scale regions and around gene promoters to identify their relationships with each other, DNA supercoiling, RNA polymerase II and sequence.

DNA supercoiling has been shown to primarily influence gene expression through increased transcription initiation (Tabuchi and Hirose, 1988). Recent analyses of several hundred gene promoters have identified a transcription and expression dependent enrichment of under-wound DNA around the transcription start site, supporting this model of gene regulation (Kouzine et al., 2013; Naughton et al.,

2013a). To identify genome-wide the distribution and regulatory potential of DNA supercoiling at gene promoters, bTMP pull-down experiments were performed and hybridised to genome wide promoter microarrays. I have analysed this data with respect to known and novel properties of gene promoters to characterise the relationship between DNA supercoiling, sequence distribution and gene expression (Chapter 5).

In addition to a role in gene expression, DNA supercoiling has been indirectly implicated in regions of genome instability called common fragile sites (CFSs). Properties of CFSs include the convergence of DNA supercoil generating replication forks, enrichment for AT-rich flexible DNA that can absorb DNA supercoils and the formation of alternative structures which are stabilised in supercoiled DNA (Gellibolian et al., 1997; Letessier et al., 2011; Lukusa and Fryns, 2008; Le Tallec et al., 2011, 2013; Zlotorynski et al., 2003). To determine experimentally if changes in DNA supercoiling are associated with expressed CFS, I have performed bTMP pull-down experiments under conditions that express FRA3B and FRA16D and hybridised these to tiling microarrays covering these CFSs (Chapter6). Through an analysis of the 'core' and 'flanking' regions of FRA3B and FRA16D fragility, I have identified novel properties of fragility at CFSs.

2. Materials and methods

2.1 Common reagents, stock solutions and buffers

Acrylamide – 30% stock solution of 29:1 acrylamide/bisacrylamide were purchased from Severn Biotech ltd.

Carnoy's fixative 75% methanol 25% acetic acid.

Chloroform:Isoamyl Alcohol comprised of chloroform and isoamyl alcohol mixed at a ratio of 24:1.

Coomassie Brilliant Blue Stain – 45% methanol, 10% glacial acetic acid and 0.05% Coomassie Blue R250.

Coomassie Brilliant Blue Destain – 10% glacial aceitic acid and 10% methanol.

DNA Loading Buffer – 5x TBE, 40% sucrose, 0.1% orange G

DNA Markers – 100 bp ladder (NEB) was dissolved at 500 µg/ml in 1x DNA Loading Buffer. 500 ng – 1 µg was normally loaded per lane.

EDTA – Ethyl diamine-tetraaceitic acid (disodium salt) was dissolved at 0.5 M in distilled water and adjusted to pH8.0 with NaOH.

Genomic Lysis Buffer – 150 mM NaCl, 10 mM EDTA pH 8.0, 0.5% SDS.

Gentle Lysis Buffer – PBS supplemented with 5 mM EDTA, 0.5% Triton X-100, 250 µM PMSF, 1 mM DTT and NaCl to a final concentration of 150 mM-1 M.

Maleic Acid Buffer – 100 mM maleic acid, 150 mM NaCl, adjusted to pH7.5 with solid NaOH.

MOPS Running Buffer – 50 mM MOPS, 50 mM Tris-HCl pH 7.6, 1 mM EDTA, 0.1% SDS.

PBS – Dublecco's PBS. Comprised 10 mM phosphate, 0.137 M NaCl, 27 mM KCl. Solution sterilised by autoclaving.

Phenol:Chloroform:Isoamyl alcohol - phenol, chloroform and isoamyl alcohol mixed at a ratio of 25:24:1.

Proteinase K was dissolved at 50 mg/ml in 50 mM Tris-HCl (pH 7.5), 2.5 mM $CaCl_2$ and 50% glycerol and stored at -20°C.

Semi-Dry Transfer Buffer – 24 mM Tris-HCl pH7.6, 192 mM glycine, 10% SDS, 20% methanol.

SDS Lysis Buffer – 50 mM Tris-HCl pH 6.8, 1% β-mercaptoethanol, 2% SDS, 0.1% bromophenol blue, 10% glycerol.

Sonication Buffer – 5 M urea, 2 M NaCl.

SDS – 10% and 20% stocks were prepared in distilled water.

Sodium Acetate – 3 M pH 5.2

SSC (20x stock) – 3 M NaCl, 0.3 M sodium citrate.

TBE Buffer (10x stock) – 500 mM Tris-borate, 1 mM EDTA.

TBS-T – 137 mM NaCl, 2.7 mM KCl, 25 mM Trsi-HCl pH 7.6, 0.1% Tween 20.

TE Buffer – 10 mM Tris-HCl pH 7.6, 1 mM EDTA.

TEEP20N – 10 mM Tris-HCl pH 7.6, 0.5 mM EDTA, 0.5 mM EGTA, 0.25 mM PMSF, 0.05% NP40, 20 mM NaCl.

TEEP80N – 10 mM Tris-HCl pH 7.6, 0.5 mM EDTA, 0.5 mM EGTA, 0.25 mM PMSF, 0.05% NP40, 80 mM NaCl.

Wet Transfer Buffer – 25 mM Tris-HCl pH 7.6, 200 mM glycine, 10% methanol.

2.2 Cell culture

The cell lines used were human retinal pigmented epithelial (RPE1) and human lymphoblastoid (SWEIG and Neo3) cells. Cells were maintained at 37°C in a humidified atmosphere containing 5% CO_2. RPE1 cells were cultured in Roswell Park Memorial Institute media (RPMI, Invitrogen) supplemented with 3 mM glutamine, 10% FCS, 100 U/ml penicillin, 100 μg/ml streptomycin, 8.1 mg/L phenol red, 2 mg/ml pyruvate, 6 mg/ml oxalacetic acid, 1x MEM non-essential amino acids (Invitrogen) and 3.75 mM MOPS. Neo3 and Sweig cells were cultured in Dubleco's modified eagle media/nutrient mixture F12 (DMEM/F12, Invitrogen) supplemented with 3 mM glutamine, 0.34% sodium bicarbonate, 10% heat-inactivated fetal calf serum, 100 U/ml penicillin, 100 μg/ml streptomycin and 8.1 mg/L phenol red. Fetal calf serum (FCS) provides basic nutrients, hormones and growth factors. All cell culture manipulations were undertaken in a laminar flow hood. To avoid bacterial and fungal contamination of the cell cultures, all objects and surfaces were sprayed with 70% ethanol before use.

2.2.1 Passaging cells

All media and solutions were pre-warmed to 37°C in a water bath before use. For adherent cells (RPE1) the culture medium was aspirated and the cells rinsed in PBS. Trypsin/Versene solution was added to cover the cells and the cultures incubated at 37°C until the cells detached. The flasks were tapped to ensure complete dissociation of cells from the surface and checked visually with a microscope. Culture medium was added to inactivate the trypsin and the cell clumps dissociated by gentle pipetting. The cells were transferred to a 15 ml Falcon tube and pelleted by centrifugation (1200 rpm, 4 minutes, room temperature in a benchtop centrifuge). The cell pellet was resuspended in 5 ml fresh culture medium and re-seeded at a suitable density.

For non-adherent cells (SWEIG, Neo3) the suspension was transferred to a 15 ml Falcon tube, dissociating cell clumps by pipetting, and pelleted by centrifugation. To

wash the cells the pellet was resuspended in 5 ml PBS and re-pelleted by centrifugation. The cell pellet was resuspended in 5 ml complete medium, cells werecounted using a Coulter Particle Count and Size Analyser, and adjusted to give a density of 300,000 cells/ml.

2.2.2 Cryopreservation and liquid nitrogen recovery

Cells were harvested as described in Section 2.2.1 and the pellet resuspended in 1 ml Freezing Media (10% DMSO, 90% fetal calf serum) per 3 million cells. The cells were frozen in 1 ml aliquots, initially at -80°C for 24 hours before being transferred to liquid nitrogen for long term storage. To recover cells from liquid nitrogen aliquots were rapidly defrosted in a 37°C water bath, added to pre-warmed media in a small flask and cultured.

2.2.3 Drug treatment

To dissect the molecular basis of RNA polymerases, topoisomerases and common fragile sites, cells were treated with a number of inhibitors. To stall topoisomerases prior to chromatin immunoprecipitation the topoisomerase I inhibitor camptothecin (5 μM) and the topoisomerase II inhibitor ICRF193 (35 μM; Biomol) were added to cells for 3 hours. To inhibit transcription α-amanitin (50 μg/ml) was added to cells for 5 hours and wash-out samples were taken 3 hours after replacing the media following α-amanitin treatment. To express common fragile sites low concentrations of the DNA polymerase inhibitor aphidicolin was added to cells for 24 hours (0.4 μM for RPE1 cells, 0.6 μM for Neo3 cells) and to arrest cells at metaphase the spindle formation inhibitor colcemid (100 ng/ml) was added to cells for 30 minutes.

2.3 DNA preparation and analysis

2.3.1 Genomic DNA preparation

Genomic DNA was typically prepared from $1x10^6$- $1x10^7$ cultured cells. Cells were harvested as described and resuspended in 500 µl PBS. An equal volume of 2x Genomic Lysis Buffer was added and incubated at room temperature for 10 minutes. The sample was treated with RNAase A/T1 (40 µg/ml RNAase A, 100 U/ml RNAaseT1) and incubated at 37°C for 30 min, followed by proteinase K (200 µg/ml) at 50°C overnight. The following day the samples were extracted twice with phenol:chloroform by adding an equal volume of 25:24:1 Phenol:Chloroform:Isoamyl Alcohol (minimum sample size 100 µl), inverting to mix, and incubating on ice for 5 min. The sample was centrifuged at 14,000 rpm for 5 min at room temperature in a bench top centrifuge. The DNA-containing aqueous phase was transferred to a fresh tube, leaving peptides in the residual phenol. To remove residual phenol an equal volume of chloroform:isoamyl alcohol was added to the sample, incubated on ice and centrifuged. DNA was precipitated by adding NaAc to 0.3 M, 2-2.5 volumes of EtOH and glycogen (1 µg/µl) as carrier, mixing thoroughly by inversion and incubating for > 30 min on dry ice or overnight at -20°C. To pellet the precipitated DNA, samples were centrifuged at 14,000 rpm for 30 min at room temperature in a benchtop centrifuge. To removed salts the pellet was washed once in 70% ethanol and centrifuged at 14,000 rpm for 5 min. The supernatant was removed and the pellet air dried, followed by re-suspension in ultra-pure water (MilliQ) or TE buffer.

2.3.2 DNA quantification

DNA sample concentration and purity was assessed using a nanodrop spectrophotometer. The A_{260} value was used to calculate the sample concentration, with $1A_{260}$ equivalent to 50 µg/ml of double stranded DNA. The spectrum of absorbance and the A_{260} /A_{280} ratio was used as an estimation of DNA sample purity.

2.3.3 Agarose gel electrophoresis

DNA preparations were size fractionated by horizontal gel electrophoresis through 1-2 % agarose gels in TBE Buffer. Buffers were supplemented with 0.5 μg/ml ethidium bromide and electrophoresed at 90 V for approximately one hour. DNA samples were loaded in 1x DNA Loading Buffer alongside 1 Kb or 100 bp standards. The DNA was imaged following electrophoresis on a UV transilluminator or laser scanner (Fuji FLA5100).

2.4 Protein preparation and analysis

2.4.1 Preparing protein extracts

Whole cell protein extracts were prepared by washing cells in PBS then lysing with 1x SDS Lysis Buffer in the plate. The cells were recovered by scraping and were transferred to a 1.5 ml Eppendorf tube, boiled at 95°C for 5 minutes to inactivate proteases and sonicated to solubilise (5 seconds, 5 μ).

2.4.2 Poly-acrylamide gel electrophoresis (SDS-PAGE)

Protein lysates were analysed Bis-Tris acrylamide gels constructed in 1.0 mm thick plastic cassettes with a twelve well comb. First a resolving gel was prepared with 7%, 10% or 12% 29:1 acrylamide/bis-acrylamide in resolving buffer (357 mM Bis-Tris pH 8.0, 1:5000 ammonium perisuphate (APS), 1:500 tetramethylethylenediamine(TEMED)). The gel solution was poured into the cassette and a water saturated butanol overlay carefully applied. Once the gel was set (30 minutes) the butanol overlay was flushed. A stacking gel was prepared (5% 29:1 acrylamide/bis-acrylamide, 357 mM Bis-Tris pH 8.0, 1:1250 APS, 1:333

TEMED) and poured on top of the resolving gel. A gel comb was placed in the top of the gel and allowed to set (5 minutes). The gel was mounted in the electrophoresis apparatus in MOPS Running Buffer. The comb was removed and the wells were flushed. Protein samples in 1x SDS Lysis Buffer and size standards were applied to the wells of the gel and electrophoresed, typically at 200 V for 1 hour. To assess protein sample quality and relative concentration between samples, proteins were stained with Coomassie brilliant blue.

2.4.3 Western blotting

To analyse specific proteins the fractionated protein samples were transferred from a gel to a polyvinylidene fluoride (PVDF) membrane by either wet transfer (> 90 KDa) (Amersham Transphor Unit) or semi-dry transfer (< 90 KDa) (BioRad Transblot SD). The PVDF membrane was soaked briefly in methanol before washing in Wet- or Semi-Dry- Transfer Buffer. The transfer apparatus was then assembled in the appropriate Transfer Buffer.

For antibody probing the membranes were blocked with 5% non-fat dry milk protein (Marvel) in TBS-T for 1 hour at room temperature. Membranes were incubated in primary antibody diluted in block for 1 hour at room temperature or overnight at 4°C. To remove unbound primary antibody membranes were washed 3x 10 minutes in TBS-T. Membranes were then incubated with a horseradish peroxidase (HRP) conjugated secondary antibody in block for 1 hour at room temperature. Membranes were again washed 3x for 10 minutes in TBS-T to remove non-specific binding. The signal was visualised using chemiluminesence (ECL, Thermo Scientific). An equal volume of each ECL solution was added to the membrane and incubated at room temperature for 1 minute. The membrane was exposed to film (Fuji) and developed.

2.4.4 Immunofluoresence

To analyse the distribution of specific proteins in cells, cells were fixed, probed with antibodies to the protein of interest and imaged by fluorescent microscopy. Adherent

cells were seeded onto sterile glass slides and grown for 24 hours, whilst non-adherent cells were cytospun onto slides following the manufacturer's instructions (Thermo Cytospin). Slides were washed gently with PBS and the cells fixed in 4% paraformaldehyde (PFA) for 7 minutes rocking at room temperature. Fixed slides were washed 3x 5 minutes in PBS and the cell/nuclear membranes permeabilised with 0.2% Triton X100 for 10 minutes. Slides were again washed 3x 5 minutes in PBS. To block non-specific binding cells were incubated in 5% horse serum (HS) for 30 minutes under parafilm in a humidified chamber. Primary antibodies were diluted 1:50 in 5% HS and applied to the slides which were incubated overnight under parafilm in a humidified chamber. The following day slides were rinsed 3x for 5 minutes in PBS and incubated for 1 hour in secondary antibody diluted 1:100 in 5% HS. Slides were mounted in Vectashield and DAPI (500 ng/ml) and sealed with rubber cement (Pang). After 1 hour the slides were imaged by fluorescent microscopy (Zeiss) at 40x or 100x magnification, using IPLab software to control the microscope.

2.5 Chromatin preparation and analysis

2.5.1 Salt extraction method

To characterise the stability of protein-DNA interactions in chromatin a salt extraction method was used. Adherent cells were seeded into 10 cm Petri dishes and grown overnight to 80% confluency. If required, the cells were first fixed at a range of formaldehyde concentrations for 10 minutes. Cells were washed in ice cold PBS and incubated in 1.2 ml Gentle Lysis Buffer supplemented with NaCl to the required final concentrations (either 150 mM, 300 mM, 500 mM, 750 mM or 1 M) for 30 minutes at 4°C. Cell lysates were transferred to a 1.5 ml Eppendorf tube with a cell scraper and centrifuged at 14,000 rpm for 15 minutes at 4°C in a benchtop centrifuge. The supernatant containing the soluble sample was transferred to a new tube and the insoluble sample was resuspended in 1.2 ml ice cold lysis buffer. A 50

µl sample of the ‘chromatin associated’ and ‘free’ protein was combined with 2x SDS Loading Buffer, incubated at 95°C for 5 minutes, sonicated and analysed by SDS-PAGE and western blotting.

2.5.2 Chromatin preparation by sucrose gradient sedimentation

2.5.2.1 Nuclei preparation

To isolate nuclei cell membranes were lysed under conditions that leave the nuclear membrane intact. Cells were harvested as described and pellets resuspended in 5 ml ice cold NBA Buffer (85 mM KCl, 10 mM Tris-HCl pH 7.6, 5.5% sucrose, 0.5 mM spermidine, 0.2 mM EDTA, 0.25 mM PMSF). To lyse the cell membranes 5 ml of ice cold NBB Buffer (85 mM KCl, 10 mM Tris-HCl pH 7.6, 5.5% sucrose, 0.5 mM spermidine, 0.2 mM EDTA, 0.25 mM PMSF, 0.2% NP40) was added and the solution incubated on ice for 3 minutes, inverting periodically to mix. Nuclei were pelleted by centrifugation at 2500 rpm for 4 minutes at 4°C. The supernatant was aspirated and the pellet resuspended in 5 ml NBR Buffer (85 mM KCl, 10 mM Tris-HCl pH 7.6, 5.5% sucrose, 0.5 mM $MgCl_2$, 1.5 mM $CaCl_2$, 0.25 mM PMSF). The nuclei were again centrifuged at 2500 rpm for 4 minutes at 4°C and the pellet resuspended in 200 µl NBR Buffer. The concentration of nuclei was determined by adding 5 µl of nuclei in NBR to 95 µl NBR Buffer and incubated with 1 µl DNaseI (100 µg/ml) for 5 minutes at room temperature. To this fragmented DNA sample 400 µl of Sonication Buffer was added and the DNA concentration measured using a spectrophotometer.

2.5.2.2 Preparation of soluble chromatin

To prepare soluble chromatin nuclei were adjusted to 20 A_{260} and 1 ml aliquots were RNase treated (10 minutes, room temperature) and the DNA briefly digested with 10 U/ml MNase at room temperature for 10 minutes. To inactive the MNase 20 mM EDTA was added and the nuclei were pelleted by centrifugation at 5000 rpm for 30

seconds. The supernatant was discarded and the nuclei resuspended in 850 µl TEEP20N Buffer and left overnight at 4°C to release soluble chromatin. Soluble chromatin was recovered by taking the supernatant following centrifugation at 13,000 rpm for 5 minutes at 4°C.

2.5.2.3 Sucrose gradient sedimentation

To purify soluble chromatin samples were centrifuged onto a 50% sucrose cushion in TEEP80N Buffer. To prepare sucrose gradients 1.5 ml 50% sucrose in TEEP80N was pipetted into an SW55 (Beckman) tube and 10% sucrose in TEEP80N layered carefully on top. A chromatin sample up to 1 ml was carefully layered on top of the gradient. The sample was centrifuged at 50,000 rpm for 50 minutes at 4°C in an ultracentrifuge. Following centrifugation samples were fractionated by upwards displacement into 12 aliquots, with the first aliquot taken from the top (i.e. lowest sucrose density). The relative concentration of DNA in each sample was measured by UV absorbance at 260 nm during the fractionation and the peak fractions containing the chromatin were stored at 4°C for further analysis.

2.5.3 Chromatin immunoprecipitation (ChIP)

2.5.3.1 Sonicated chromatin preparations

To investigate the distribution of proteins on DNA in cells the ChIP technique was used. For each condition cells were seeded in duplicate into 10 cm Petri dishes and grown to 80% confluency. To stabilise protein-DNA interactions cells were cross-linked with 0.5% formaldehyde for 10 minutes at room temperature. To quench the unreacted formaldehyde 100 mM glycine was added and the cells incubated for 5 minutes. Cells were washed twice in ice cold PBS and transferred to a 1.5 ml Eppendorf tube using a cell scraper. Cells were pelleted by centrifugation at 2,000 rpm for 4 minutes at 4°C. The supernatant was removed and the pellets were resuspended in 200 µl ChIP Lysis Buffer. Duplicate plates were combined to give a total of 400 µl and the chromatin fragmented by 13 rounds of 30 seconds on/off

sonication at 2 μ on ice (Soniprep 150 probe sonicator, MSE). Following sonication cell debris was recovered by centrifugation at 14,000 rpm for 15 minutes at 15°C and the supernatant divided into two 200 μl aliquots as technical replicates. Chromatin samples were used immediately or stored at -20°C.

2.5.3.2 Immunoprecipitation

Sonicated chromatin samples were diluted in 1300 μl ChIP Dilution Buffer and a 50 μl aliquot taken as 'input'. Chromatin samples were pre-cleared for 1 hour by rotating at 4°C with rabbit immunoglobulins and immunogen coated beads appropriate for the specific primary antibody used and blocked with salmon sperm DNA. The beads were separated from the supernatant using a magnetic rack (Dynabeads) or centrifugation at 2000 rpm for 2 minutes (agarose beads) and the pre-cleared chromatin supernatant was transferred to a new 2ml tube. To immunoprecipitate a specific protein of interest, 50 μl of primary antibody and 50 μl of appropriate beads were added to the chromatin sample and incubated at 4°C overnight on a wheel.

The following day non-specific protein associations were washed from the beads by successive 5 minute incubations at 4°C in TSEI, TSEII, Buffer 3 and twice in TE. Bound chromatin was eluted by incubating the beads in 250 μl of ChIP Elution Buffer for 30 minutes at room temperature. The beads were separated from the supernatant and transferred to another 1.5 ml Eppendorf tube. A second elution from the beads in 200 μl ChIP Elution Buffer for 15 minutes was combined with the first, giving the immunoprecipitated chromatin sample.

2.5.3.3 DNA purification and microarray hybridisation

To reverse the DNA-protein formaldehyde cross-links samples and inputs were incubated in 200 mM NaCl for 6 hours at 65°C. To purify the DNA samples were treated with proteinase K (100 μg/ml) for 1 hour at 55°C and purified using a Qiagen MinElute PCR purification kit, following the manufacturer's instructions and eluting

in 10 µl of PCR grade water. Input DNA concentrations were measure at 260 nm and diluted to 3 ng/µl. Samples and 10 µl of diluted inputs were whole genome amplified (Sigma GenomePlex) following the manufacturer's instructions. Amplified DNA samples were quantified and visualised by agarose gel electrophoresis. If necessary the samples were re-amplified (Sigma GenomePlex re-amplification). Amplified ChIP samples and inputs were sent to the VUMC microarray facility for random prime labelling with Cy3 or Cy5 (ENZO) and microarray hybridisation. Following hybridisation and washing following the manufacturer's instructions slides were scanned on an Agilent Microarray scanner at 2 micron resolution, generating a TIFF file for analysis by Agilent feature extraction software. This software produces a PAIR file of probe intensities, which forms the basis of all subseqent analysis.

2.5.4 Chromosome analysis

To analyse chromosome structure and genome stability, brightfield microscopy on Giemsa stained metaphase chromosomes was performed under conditions that induce common fragile sites. Cells were seeded into T25 flasks and allowed to settle. To induce common fragile site expression the DNA polymerase inhibitor aphidicolin was added to cells. Following 24 hours aphidicolin treated and control cells were harvested as described. Cell pellets were resuspended in 5 ml PBS to wash and centrifuged at 1200 rpm for 4 minutes at room temperature to re-pellet cells. To swell the cells the cell pellet was resuspended in 75 mM KCl applied by pastette with gentle vortexing up to a volume of 10 ml and incubated for 10 minutes at room temperature. The swelled cells were collected by centrifugation at 1200 rpm for 4 minutes at room temperature and resuspended in Carnoy's fixative with gentle vortexing up to a volume of 10 ml and incubated for 10 minutes at room temperature. These fixed cells were re-pelleted by centrifugation and resuspended in Carnoy's fixative twice more and stored at -20°C overnight before use.

Following overnight incubation samples were warmed by hand, pelleted by centrifugation and resuspended in an appropriate amount of fresh Carnoy's buffer. Microscope slides were stored in ethanol with a drop of HCl to give micro-scratches

on the slide surface. Slides were removed from the ethanol, excess ethanol removed with tissue and air dried. To drop the fixed metaphase preparations the slide was humidified by gentle blowing and a small volume of fixed cells dropped by pastette from a height of 15 cm and dispersed with a gentle blow. Slides were allowed to air dry and the success of a drop was assessed by phase contrast microscopy. Cell concentration and droplet size were then adjusted to get the optimum metaphase conditions. Once conditions were optimised, air dried slides were baked at 65°C for 1 hour or left on the bench for several days. Slides were then stored at room temperature until use.

To identify chromosomal aberrations in metaphase chromosomes, samples were karyotyped by Giemsa banding. Slides were rehydrated in PBS for 2 minutes and trypsin treated for 1 minute to improve subsequent staining. Trypsin was inactivated with PBS+$MgCl_2$ and the slides were stained for 13 minutes in 6% Giemsa rocking at room temperature. Slides were washed 3x 5 minutes in tap water, air dried and fixed/mounted in DEPEX mounting medium for imaging by brightfield microscopy.

2.6 bTMP analysis of DNA supercoiling

2.6.1 bTMP synthesis

BTMP was synthesised by Nicolaos Avlonitis of the University of Edinburgh Chemistry department to produce the molecule first described in Saffran et al. (1988). The synthesised molecule was stored in powdered form at -20°C, with a working aliquot stored in methanol at -20°C.

2.6.2 bTMP sequence specificity

2.6.2.1 bTMP oligonucleotide photo-crosslinking

To identify whether bTMP preferentially binds GC or AT dinucleotides a photocrosslinking experiment was performed on poly-GC and poly-AT oligonucleotides (Sigma). Furthermore, to identify whether bTMP preferentially binds A form, B form or A/B intermediate form DNA helical structures, a photo-crosslinking experiment was performed on oligonucleotides that form these structures (Table 2.1) (Hays et al., 2005). 5 µg of oligonucleotides and 30 ng of bTMP were diluted to 100 µl in TE Buffer and incubated in one well of a 96 well plate for 10 minutes in the dark at room temperature, alongside no DNA and no bTMP controls. To photo-crosslink the bTMP to the DNA, samples were incubated for 15 minutes in 320-400 nm UV light. Non cross-linked control samples were further incubated in dark for 15 minutes. DNA oligonucleotides were purified using a QiaQuick PCR clean up kit (Qiagen) and eluted in 30 µl PCR quality water (MilliQ). DNA samples were prepared for dotblot by adding 1 µl of 20x SSC to 9 µl of purified DNA and ananlysed as described in section .6.4.

Type	Sequence
A form	CCTCCGGAGGCCTCCGGAGGCCTCCGGAGG
B form	CCTGCGCAGGCCTGCGCAGGCCTGCGCAGG
AB form	CATGGGCCCATGCATGGGCCCATGCATGGGCCCATG

Table 2.1 Sequences of alternative DNA helix oligonucleotides.

2.6.2.2 bTMP photo-crosslinking mass spectrometry

To identify the sequence bias of bTMP photo-crosslinking 5 µl of sonicated genomic DNA (300-500 bp) was combined with 3 µg of psoralen in a total of 60 µl TE and incubated in the dark at room temperature for 10 minutes, followed by 15 minutes in 320–400 nm UV light. Non-crosslinked control samples were incubated for a total of 25 minutes in the dark. DNA was purified from unbound psoralen using a G50 sephadex column and diluted to 20 ng/ul. A 50 µl aliquot of psoralen-bound and control DNA was incubated in RNaseAT1 for 4 hours at 37°C and the DNA purified from ribonucleotides in a G50 sephadex column. These DNA samples were digested to mononulotides in Mononucelotide Digestion Mix for 6 hours at 37°C. Photo-crosslinked samples and non-crosslinked controls were analysed by HPLC-MS to identify bTMP bound nucelotides, in conjunction with the Chemistry department.

2.6.3 bTMP in cell DNA photo-crosslinking

To analyse DNA supercoiling bTMP molecules were photo-crosslinked in living cells. Cells were seeded in duplicate in 10 cm Petri dishes and grown to 80% confluency overnight. The following day cells were washed 3x in PBS followed by bTMP at 500 ng/ml (RPE1) or 1.4 µg/ml (Neo3) in a total volume of 1 ml PBS under a parafilm disk, to ensure an even distribution of drug across the plate. Cells were incubated in bTMP for 20 minutes in the dark followed by 10 minutes photo-crosslinking in 320-400 nm UV light. Cells were washed 3x in PBS and scraped into a 1.5 ml Eppendorf tube in 1 ml PBS. Samples were centrifuged at 2000 rpm for 4 minutes at 4°C, the supernatant discarded and resuspended in 200 µl ChIP Lysis Buffer. Duplicate plates were combined to give 400 µl total. bTMP cross-linked samples were fragmented by 13 rounds of 30 seconds on/off sonication at 2 µ on ice. Following sonication samples were incubated in 10 µg/ml proteinase K for 4 hours or overnight at 55°C. DNA was isolated by phenol:chloroform extraction and resuspended in 50 µl TE buffer to give the bTMP photo-crosslinked DNA sample.

2.6.4 bTMP DNA dotblot

To confirm photo-crosslinking of bTMP into DNA in cells, 1 µl of photo-crosslinked DNA sample was diluted to give 10 µl in 1x SSC Buffer and applied to a charged nitrocellulose membrane (GE healthcare) 2 µl at a time. After each 2 µl application the membrane was allowed to air dry before repeat applications to the same spot. A serial dilution of biotinylated-oligonucleotides was included as a biotin-standard. Following sample application the DNA was UV crosslinked to the membrane at 150 mJ/cm^2. The membrane was blocked in 1x Blocking Reagent (Roche) diluted in Maleic Acid Buffer for 30 minutes rocking at room temperature. Following blocking the membrane was incubated in avidin-conjugated HRP antibody in 1x Blocking Reagent for 1 hour rocking at room temperature or 4°C overnight. The membrane was washed twice in Maleic Acid Buffer and twice in TBS-T for 5 minutes rocking at room temperature. The signal was visualised using chemiluminesence (ECL, Thermo Scientific) and exposed to X ray film (Fuji) and developed.

2.6.5 bTMP immunoprecipitation

To enrich for the under-wound DNA preferentially bound by bTMP immunoprecipitation for the biotin tag was carried out on bTMP photo-crosslinked DNA samples. Samples were divided into two technical replicates, with one stored at -20°C. The samples for immunoprecipitation were diluted in 900 µl TE Buffer and 50 µl taken as 'input'. To each sample 50 µl of pre-washed streptavidin coated magnetic beads were added and incubated for 2 hours at room temperature then 4°C overnight on a wheel. The following day non-specific DNA was washed from the beads by 5 minute washes at 4°C in buffers TSEI, TSEII, Buffer 3 and twice in TE on a wheel. Bound DNA samples and were eluted from the streptavidin beads in 50 µl Biotin Elution Buffer at 90°C for 10 minutes and the immunoprecipitated DNA sample transferred to a new tube. The samples were made up to 200 µl with PCR quality water and the DNA purified with a Qiagen MinElute PCR purification kit following the manufacturer's instructions and eluting in 10 µl PCR grade water. Samples were amplified and the DNA quantitified by spectrophotometer and agarose

gel electrophoresis, before being sent to the VUMC microarray facility for labelling and microarray hybridisation.

2.7 Microarray Design

To identify the distribution of topoisomerases and changes in DNA supercoiling at common fragile sites, microarrays were custom designed to cover regions of the genome with distinct chromatin features; including gene rich, gene poor, common fragile sites and telomeric loci (Table 2.2). These 180k probe arrays were designed by Agilent to have unique 50 bp probes with a coverage of 1 probe per 60 bp. Previous studies in the lab had used these custom arrays to determine DNA supercoil domains by psoralen-IP and the distribution of RNA polymerase II (Naughton et al., 2013). This was directly comparable with my data allowing a thorough investigation of the interplay between DNA structure, transcription and topoisomerase activity at these loci. For a more comprehensive analysis of topoisomerases at gene promoters, ChIP samples were hybridised to Nimblegen chromosome 11 tiling arrays, which contain 2,509 transcription start sites. To investigate the distribution of DNA supercoiling at promoters genome-wide by psoralen-IP, samples were hybridised to Nimblegen 2.1M promoter arrays (name). These microarrays cover regions 7 kb upstream to 2 kb downstream of each promoter in the human genome, with a coverage of 1 probe per 100 bp.

Region	*Size (Mb)*	*Probes*	*Spacing*	*Chr*	*Start*	*End*
IGBP1	2	11047	60	X	68369744	70369744
LDHA	2	15176	60	11	17417960	19417960
11p15.5	2.8	26126	60	11	1	2800000
Enr312	0.7	6965	60	11	131031152	131732236
Enr332	0.6	5501	60	11	64120923	64720922
Xq25	3.9	21830	60	X	119145001	123045000
RNU2	0.5	3259	60	17	41124772	41624771
RNU1	0.55	4447	60	1	16570000	17120000
11p13	5.5	43382	60	11	27100001	32600000
FRA3B	5.1	26807	100	3	58600001	63700000
FRA16D	2.5	15305	91	16	77512501	80012500

Table 2.2 Custom tiling microarray design.

2.8 Bioinformatic analysis

To analyse the microarray data generated by ChIP-chip and psoralen-IP experiment and relate these results to the work of others a bioinformatic approach was employed. Unless otherwise stated, all analysis was performed in the R language (http://www.r-project.org/) (version 2.15.1) and associated packages, in particular those of Bioconductor (http://www.bioconductor.org/). Other analysis was performed in Perl, Galaxy (http://galaxyproject.org/) and using UCSC binary utilities through UNIX (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/).

2.8.1 Datasets

Novel tiling array datasets were generated by ChIP-chip in RPE1 cells for topoisomerase I, topoisomerase IIα and topoisomerase IIβ and by psoralen-IP in

NEO3 cells under control, α-amanitin, α-amanitin reverse, aphidicolin, bleomycin and genomic DNA conditions. Novel promoter array datasets were generated for psoralen-IP samples under control, α-amanitin, α-amanitin reverse and genomic DNA conditions.

Other datasets in the lab were compared with the custom tiling, chromosome 11 and promoter array data. These included RPE1 datasets using the same custom microarray platform for RNA-polymerase ChIP-chip, cDNA and psoralen-IP samples under control, α-amanitin, α-amanitin reverse, bleomycin and genomic DNA conditions. Genome wide expression array data (Illumina HT12 array) for RPE1 cells was also generated within the lab. Expressed genes were called based on the top quartile of expression values and non-expressed genes on the bottom quartile of expression values.

External datasets were used extensively including the ENCODE project (http://genome.ucsc.edu/ENCODE/) and BioMart database (http://www.biomart.org/biomart/martview/) to identify datasets including gene positions, CpG islands positions, transcription factor binding sites, structural protein binding sites, etc.

2.8.2 General bioinformatic analyses

2.8.2.1 Parametric and non-parametric statistics

The correct use of parametric and non-parametic statistics give increased statistical power. Parametric methods make assumptions about the data, such as the expectation that the data is normally distributed, which should be met in order for the analysis to be considered robust. In some cases, I have displayed or analysed data using a parametric test where a non-parametric test would have been more appropriate (e.g. used a t test rather than Mann-Whitney U). In each case tested, a re-calculation using either parametric or non-parametric statistics gave essentially the same value.

2.8.2.2 Correlation

To identify whether there is a relationship between two samples the Pearson's product-moment correlation coefficient was performed using the 'cor.test' function. The Pearson's correlation calculates the covariance of the two samples and divides this by the product of their standard deviations. The correlation between the two samples is expressed as a value between 1 (perfect correlation), 0 (no correlation) and -1 (perfect anti-correlation). The probability that this correlation represents a significant deviation from chance is given as a p-value.

2.8.2.3 Student's t test

A two-tailed two-sample t test identifies whether the difference between two samples is significant by comparing sample population means. The 't.test' command was used and the significance determined from the assigned p-value.

2.8.2.4 Data distribution

A number of commands were used to determine the distribution of data both numerically and graphically. The 'summary' and 'boxplot' commands were used to represent the median, interquartile range and outliers of datasets either numerically or graphically. In most cases the boxplot outliers were not informative and were removed from figures ('outline=F'). To represent the distribution of a single dataset in a continuous manner, histograms were plotted with the 'hist' command. Other types of graph were produced with the 'plot' command, including line graphs ('type=l') and scatter plots ('type=p'). To represent the density of data within a scatter graph, which was often obscured due to the high number of datapoints, the 'smoothScatter' command was used. Venn diagrams were plotted using the 'venn' command in the 'gplots' package.

2.8.2.5 Two colour microarray data processing and analysis

Microarray scans were provided by the VUMC microarray facility as TIFF images and PAIR raw intensity files. Each TIFF image was inspected for scratches, bubbles and other hybridisation artefacts before subsequent analysis. Signal intensity files were read into R using the Ringo Bioconductor package designed for the analysis of two-colour oligonucleotide arrays (Toedling, 2012; Toedling et al., 2007). The relative distribution of signal intensity for each fluorophore was established within and between arrays using the 'plotDensities' command and the signal intensity bias identified using 'ma.plot'. Normalisation of signal differences within and between arrays was performed in an experiment dependent manner using either Variant Stabling Normalisation (VSN) or a combined loess/scale normalisation in the 'Limma' Bioconductor package. The post-normalisation data was checked by 'plotDesities' and 'ma.plot' to ensure a more similar relative distribution and no signal bias. Once satisified the probe signal values were combined with their corresponding genomic position for further analysis in R and written to BED files for analysis in the UCSC genome browser and Galaxy.

2.8.2.6 Custom track analysis with the UCSC genome browser

To analyse the distribution of custom datasets with respect to genomic position and a vast array of biological information, the normalised microarray data was uploaded to the UCSC genome browser. Smaller files were uploaded directly as a BED file with a track header used to build the image, for example:

track type=bedGraph name="my bed" description="a graph of my bed" visibility=full color=200,100,0 altColor=0,100,200 priority=20 graphType=bar autoScale=off

Larger datasets were converted to the bigwig file format using the 'bedGraphToBigWig' programme and uploaded via an FTP server. In this case only

a portion of the data required is stored on the UCSC server, with the full dataset remaining on the local server. To load bigwig data a track header was pasted into the 'Paste URLs or data' box of the 'Add Custom Track' window, for example:

track type=bigwig name="my bigwig" description="a graph of my bed" bigDataUrl=ftp://ecrc##.med.ed.ac.uk/test_mydata.bw

2.8.3 Distribution of data around gene promoters

To directly compare the distribution of data for each promoter on the array a number of analysis tools (e.g. k-means clustering) require a matrix with the distribution organised into a set number of columns. To produce this, the distribution of data around each transcription start site was binned into 400 bp bins overlapping by 200 bp in the region 7 kb upstream to 2.8 kb downstream. This data matrix was produced by taking the median value of the probes within each bin for each TSS. If a bin contains no probes a value was imputed using the 'na.approx' command in the 'zoo' package, with the exception of the first or last bin which was replaced with the mean value for the promoter distribution. The imputation was performed up to a maximum of 3 missing values in a row and required 39 of the 49 bins to contain values. Any transcription start site which did not meet these criteria was discarded. Subsequent analysis of data at gene promoters used data in this matrix format, for the identification of median distributions around gene promoters, classification by kmeans analysis, etc.

2.8.4 Data smoothing by rolling median

To present the patterns of topoisomerase ChIP enrichment in Zoo plots and in the UCSC genome browser a 27 probe rolling median was used to clarify the patterns of data enrichment. Data smoothing was achieved using the 'rollMedian' function in the 'zoo' package. Subsequent analysis was performed on non-smoothed data, with the smoothing used solely to clarify the typical topoisomerase distribution in noisy data.

2.8.5 Determining topoisomerase domains with an edge filter

To identify the boundaries of domains of topoisomerase enrichment and depletion a rolling window edge filter was used to identify positions at which topoisomerase enrichment/depletion changes markedly. The mean values for 200 (topoisomerase I) or 400 (topoisomerase IIα/IIβ) probe windows up- and down- stream of a particular position were compared and a stringent difference cut-off determined to establish boundaries (arbitrarily defined topoisomerase I cut off 0.15, topoisomerase II cut off 0.1). Boundaries were compared with the raw signal to ensure a high degree of similarity. This analysis identifies domains successfully, with the caveats that smaller peaks of enrichment, regions with low probe density and variable regions are not easily categorised. The mean signal value was determined between each set of boundaries and enriched domains identified as those domains with a positive mean value.

2.8.6 Determining distributions around transcription start sites

Topoisomerase, RNA polymerase and DNA supercoil data were analysed relative to the transcription start site by producing matrices as outlined in Section 2.8.3. The datasets were separated into expressed and non-expressed genes and the relative distributions compared by taking the median value for each distribution around the TSS. The difference between distributions was identified by performing a t-test that compared the distributions of data in set bins. To identify enrichment at random positions on the chromosome 11 array, and therefore determine that enrichment at the promoter is genuine, a number of random positions equal to the number of expressed TSSs were chosen on the chromosome and the median distribution of topoisomerase I and IIβ determined. T-tests were performed between promoter and random data and the iteration repeated 30 times. If the p-value of the random

iteractions was lower than that of the promoter in 2 or more of the 30 iteractions, then the difference was not significant. Otherwise significance was determined based on the highest p-value attained by the random iterations (see Figure 3.17).

2.8.7 Classifying promoters based on topoisomerase distribution

To classify promoters based on topoisomerase distribution in an independent manner, kmeans clustering analysis was applied using the 'kmeans' function. Setting the cluster number to 2 identified median topoisomerase distributions with a distinct peak and a distinct trough, providing a useful segregation of the data.

2.8.8 Normalising the bTMP distribution of control/α-amanitin/wash-out data for genomic DNA

The microarray hybridisations for control/α-amanitin/wash-out samples and genomic DNA samples were performed on HG18 and HG19 Nimblegen array platforms respectively. The different array platforms have a complete probe re-design and were not directly comparable. To compare the two, matrices of promoter distribution were produced separately for each data type. The coordinates of the HG18 TSSs were converted to the corresponding HG19 TSSs and the centre point of each bin for each TSS identified as the 'probe position'. The data was subsequently rearranged into a dataframe containing the 'chromosome', 'probe position' and data for 'control', 'α-amanitin', 'wash-out' and 'genomic'. A scale normalisation was performed to ensure the control/α-amanitin/wash-out and genomic DNA datasets have the same signal distribution. Once normalised, the genomic data was subtracted from the control, α-amanitin and wash-out samples individually and the data returned to a matrix format, with rows representing TSSs and columns representing the binned data around the TSS. These datasets, normalised for genomic DNA bTMP enrichment, were used for all subsequent analysis.

2.8.9 Ranking promoters on CpG island probability

In order to rank promoters based on their probability of containing a CpG island sequence pattern, the hidden Markov model (HMM) based package 'makeCGI' was used (Irizarry et al., 2009; Wu et al., 2010). This package determines the probability of elevated CpG and GC content compared to the local sequence context, rather than the more simplistic model used by the UCSC genome browser (Gardiner-Garden and Frommer, 1987). By running the package with a low posterior probability threshold (0.5 for CpG, 0 for GC content) the majority of promoters identify some form of CpG island motif. Taking the mean probability identified for motifs within the region 500 bp up-stream to 500 bp down-stream of the TSS, where most promoter associated CpG islands are located, gives a parameter on which the promoters can be ranked based on CpG likeness. A small minority of promoters that did not contain any CpG-like motif within 1 kb centred on the TSS were added to the bottom of the distribution in an un-ordered fashion. Previous studies have identified that a posterior probability threshold of 0.99 gives the canonical CpG island list used by the UCSC genome browser (Irizarry et al., 2009) identified by a dotted line in heatmap and inflection plots.

2.8.10 Heatmap analysis of bTMP distribution

To compare the distributions of bTMP at all promoters, ordered on their probability of containing a CpG island, the data was presented as a heatmap using the 'heatmap.2' function in the 'gplots' package. To see the distribution of data clearly, the extremes of the distribution were not plotted on the heatmap instead focussing on the range of signal intensity from 5% to 95%.

2.8.11 Inflection plot for bTMP distribution

To identify the inflections in the data, where either a peak or trough is observed, the 'turningpoints' function in the 'pastecs' package was used. To limit the noise in the data, TSSs ordered on the CpG island likeness were binned into 100 TSS bins and

the median distribution for the bin determined. This distribution was smoothed with a slight loess smoothing (span=0.2), using the 'loess' function, and the inflections determined using 'turningpoints'. The inflections were plotted for each of the 206 bins to give their relative distribution with respect to TSSs.

2.8.12 Determining GC% matrix for promoters genome-wide

The comparison of sequence distribution with bTMP distribution necessitated a matrix of GC% in the bins around each TSS. To create this matrix the chromosome, bin start and bin end positions were identified for each bin of each transcription start site. This information was then used in the 'nuc' option of BEDtools (Quinlan and Hall, 2010) to determine the GC% for each bin and the matrix re-assembled so that rows represent TSSs and columns represent the binned data around the TSS.

2.8.13 Distribution of bTMP around protein binding sites

The distribution of bTMP around transcription factor and structural protein binding sites was determined for several proteins with ChIP-chip data available in the A549 epithelial cell lines through the ENCODE project. A meta-analysis of the relative enrichment of bTMP with distance from the nearest protein binding site was identified using modified in-house Perl scripts and plotted in R.

3. Mapping DNA topoisomerase I, IIα and IIβ by chromatin immunoprecipitation

3.1 Introduction

The presence or absence of specific proteins or modifications at a DNA sequence can be used to determine or predict biological function *in vivo*. For example, focal enrichment of the DNA binding protein P-300 generally identifies active enhancers (Visel et al., 2009), whereas focal enrichment for ERα at gene promoters identifies the potential for estrogen regulation (Fullwood et al., 2009) and a peak of H3K4me3 generally identifies active gene promoters (Kouzarides, 2007). Other proteins show a more local enrichment that indicates genome regulation on a larger scale, including repressive chromatin proteins (e.g. HP1α, lamin-B1) and modifications (e.g. H3K9me2/me3, H3K27me3) (Ernst et al., 2011; Guelen et al., 2008). The mapping of each of these factors by chromatin immunoprecipitation (ChIP) or DNA adenine methyltransferase identification (DamID) has transformed our understanding of gene regulation and allowed the functional annotation of eukaryotic genomes. This annotation is being performed most comprehensively by the ENCODE consortium, which has mapped hundreds of proteins/modifications in a broad range of human cell types (Dunham et al., 2012). However, the catalogue of proteins and modifications investigated through the ENCODE project is limited, with numerous important factors involved in gene regulation and genome stability still unmapped. For example, the distribution of topoisomerase enzymes in the human genome has not been identified at high resolution, despite their very high abundance and critical function in genome organisation and regulation. Many questions remain about topoisomerase function *in vivo*, including the distribution of these proteins with respect to genes, domains and one another.

3.1.1 Topoisomerase distribution in model systems

Topoisomerases are found in all forms of cellular life to relieve DNA supercoils and drive chromosome decatenation (Section 1.2.4.2). The distribution of topoisomerases in the human genome is known only for a handful of genes and at low resolution, but the conserved function of these proteins between species makes it likely that they have a conserved distribution with respect to genes and other genomic features. In addition, differences in genome organisation may strongly influence the dissipation of DNA supercoils and the corresponding difference in topoisomerase distribution. For example, the yeast genome is generally more expressed than the human genome, due to less non-coding DNA and fewer non-expressed genes. It may therefore be expected that DNA supercoiling is a more pervasive force in yeast and that topoisomerase distribution somehow reflects this. In the following section the likely distribution of topoisomerases in humans will be discussed based on a comparison of genome structure and topoisomerase distribution maps from yeast, fruit fly and non-human mammalian systems.

The yeast genome is distinct from the human genome in ways which may directly influence DNA supercoiling, including an average gene length of ~1.6 kb (compared to 10-15 kb in human), a high gene density with only ~27% of the genome non-coding (compared with 97% in human) and fewer non-expressed genes (Roger et al., 2010). Therefore, there is the potential for high levels of DNA supercoiling to transmit through the DNA to influence neighbouring genes that are in close spatial proximity. Mapping topoisomerases by ChIP hybridised to tiling microarrays (ChIP-chip) identifies that topoisomerase I and II are enriched in intergenic regions and particularly at regions flanking genes (Durand-Dubief et al., 2010). (Durand-Dubief et al., 2010). The distribution of topoisomerases within gene bodies is lower, and in the case of very highly expressed genes topoisomerase II binding is abolished altogether. This suggests a model whereby DNA supercoiling generated within short yeast genes is relieved by topoisomerase I and II in inter-genic regions.

The fruit fly genome is much more similar to that of human than yeast, with an average gene length of ~11.3 kb, low gene density with ~82% of the genome non-coding and the highly variable gene expression pattern common to multicellular

organisms (Roger et al., 2010). Interestingly, the distribution of topoisomerase I is distinct from that observed in yeast and is primarily enriched in the gene body of expressed genes (Gilmour et al., 1986). The genome wide distribution of topoisomerase I was mapped by DamID, a technique that marks DNA with adenine methylation when in contact with the protein of interest, supports an association with transcriptionally active euchromatin (Filion et al., 2010). In further contrast to the yeast model, topoisomerase II distribution is distinct from topoisomerase I in the fly genome. Mapping of topoisomerase IIβ to a small number of sites identified enrichment at AT-rich nuclear attachment regions and at DNAaseI hypersensitive sites, in a transcription dependent manner (Käs and Laemmli, 1992). A larger scale mapping of topoisomerase II cleavage patterns on genomic DNA *in vitro* also identifies an AT-rich preference, with cleavage sites extending over domains > 10 kb in length (Miassod et al., 1997). This analysis could not determine a clear consensus binding sequence and they conclude that topoisomerase II preferentially binds DNA with intrinsic curvature. Whether this distribution holds true *in vivo* is unknown, and several studies have identified significant differences between *in vitro* and *in vivo* (chromatin associated) topoisomerase cleavage sites (Käs and Laemmli, 1992; Udvardy and Schedl, 1991). Together this data supports and association between topoisomerase I and active regions of the chromatin and indicates a role for topoisomerase II at both transcriptionally repressed regions and at DNAaseI hypersensitive sites in fly. Based on the similarity of genome structure between human and fly, the fly distribution may represent a better model for the expected distribution of topoisomerase.

The genome organisation of mouse and rat are highly similar to that of the human genome and would represent the best model for the expected distribution of topoisomerases in human. However, the *in vivo* distribution of topoisomerase I has not been mapped in non-human mammalian systems. On the other hand, topoisomerase II enzymes have been mapped to the promoters of several genes in mouse, with one study identifying topoisomerase IIβ at the promoter of five out of seven genes whose expression has been shown to be dependent on topoisomerase IIβ in knock-out experiments (Lyu et al., 2006). A more detailed analysis of one of the genes identifies an enrichment for topoisomerase IIβ through the *Kcnd2* gene,

particularly at the promoter, but not upstream of the transcription start site (Lyu et al., 2006). A large scale map of topoisomerase IIβ activity in rat neuronal genes identifies enriched activity in AT-rich regions and at TSSs, including expressed and non-expressed genes (Sano et al., 2008). Therefore, the distribution of topoisomerase IIβ at gene promoters and the identification of a minority of genes regulated by topoisomerase IIβ activity supports an important role for this enzyme at promoters in mammalian cells.

The *in vivo* distribution of topoisomerases in human cells is very poorly characterised. Topoisomerase I ChIP with ten sets of primers over a 5 kb locus identifies a subtle increase in topoisomerase binding over and downstream of the expressed H2A and H2B genes, whereas no such enrichment is observed at α-satellite DNA (Khobta et al., 2006). A similar small-scale mapping experiment of topoisomerase IIα and IIβ by ChIP identifies a subtle enrichment of each at the promoter of the MLL gene (Cowell et al., 2012). To test the relationship between expression and topoisomerase binding Kouzine et al. (2013) performed ChIP for topoisomerase I and IIβ and identified their enrichment at a single position for 15 gene promoters. They identified a subtle increase in topoisomerase I binding with expression and a stronger increase in topoisomerase IIβ binding at highly expressed genes. This lead them to propose a model whereby topoisomerase I acts in a diffuse manner upstream of a transcription start site, while in addition topoisomerase IIβ is focally enriched at the transcription start site in highly expressed genes.

Based on the distributions of topoisomerases in non-human model organisms, and limited mapping data in human, several general properties of topoisomerase I and II have been observed. In general, the distribution of topoisomerase I associates with transcription and topoisomerase II with AT-rich regions and transcriptionally active gene promoters. To identify if these patterns are true for human chromatin it is essential to map topoisomerases at high resolution across a significant proportion of the genome. The aim of this chapter is to map topoisomerase I, IIα and IIβ by chromatin immunoprecipitation hybridised to microarrays (ChIP-chip) to identify topoisomerase distribution with respect to sequence, RNA polymerase, DNA structure and other parameters. In addition, a detailed analysis of topoisomerase I

and II distribution at the transcription start sites (TSSs) of human genes will be used to test the model proposed by Kouzine et al. (2013). This will provide the first comprehensive understanding of topoisomerase I and II localisation in human cells and can be used as a framework for better understanding the function of topoisomerase I and II *in vivo.*

3.1.2 Chromatin immunoprecipitation (ChIP) as a tool for mapping DNA binding proteins *in vivo*.

Chromatin immunoprecipitation identifies the DNA sequences bound by a protein of interest in living cells (Lee et al., 2006). Depending on the strength of association between the protein of interest and DNA ChIP may be performed under 'native' conditions or following a chemical cross-linking step which stably associated protein to DNA (Figure 3.1a). To isolate the chromatin for immunoprecipitation cell and nuclear membranes are lysed (Figure 3.1b) and the chromatin is fragmented to 300-500 bp by MNase digestion (native ChIP) or sonication (cross-link ChIP) (Figure 3.1c). For immunoprecipitation the chromatin extract is incubated with an antibody to the protein of interest (Figure 3.1d) which is enriched for by immunoprecipitation with an appropriate immunogenic-protein coated bead (Figure 3.1e). The DNA associated with the protein of interest is isolated from the immunoprecipitated chromatin (Figure 3.1f) and analysed by PCR (ChIP-PCR), microarray (ChIP-chip) or next generation sequencing (ChIP-seq) (Figure 3.1g). Bioinformatic analysis of microarray or next generation sequencing data can identify relative enrichments of the protein of interest at high resolution across some/all of the genome and this data can be mined to inform and answer biological hypotheses. This general protocol has been used widely to map many factors across the genome, including the many proteins/modifications mapped by the ENCODE consortium (Dunham et al., 2012).

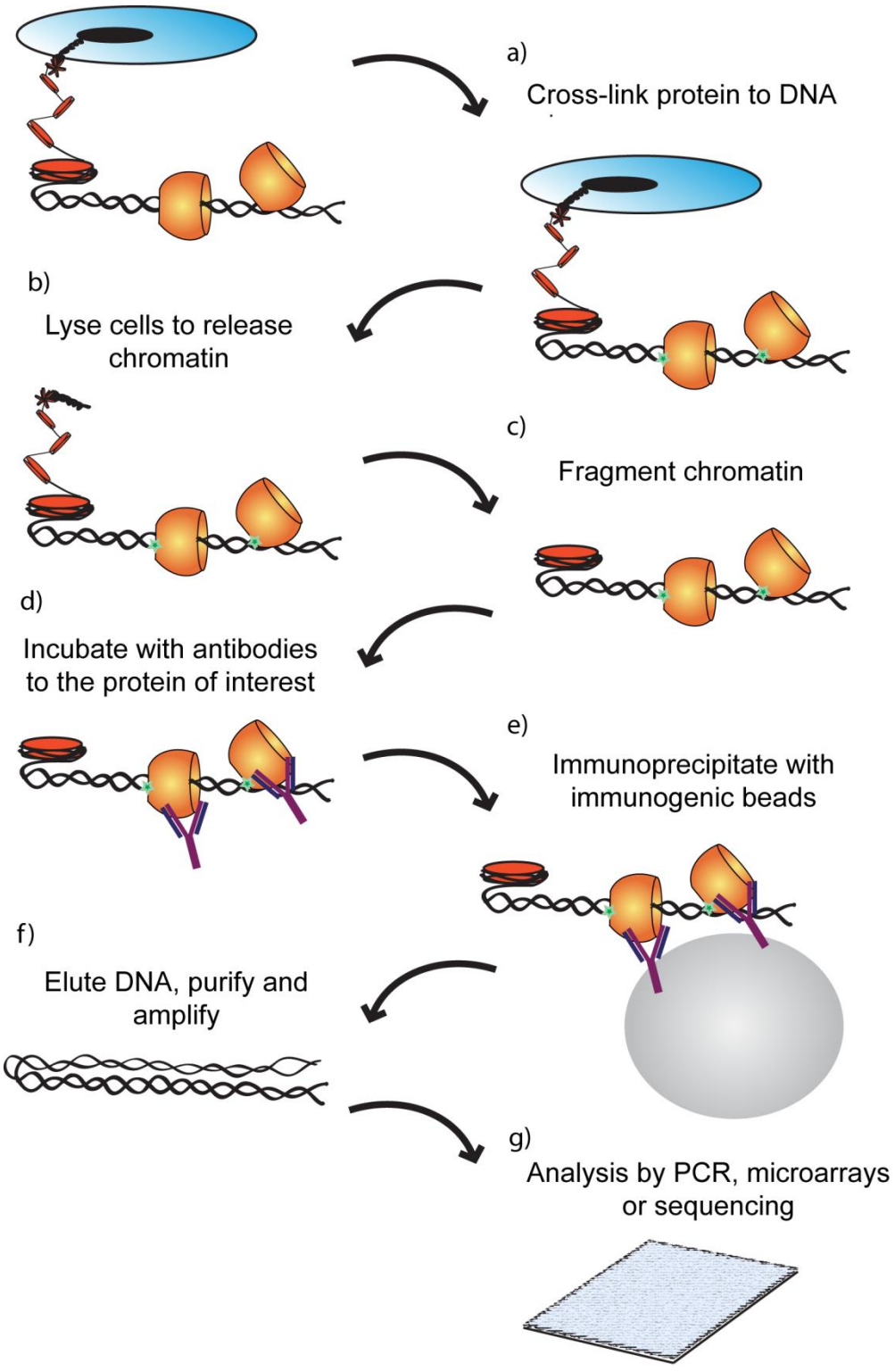


Figure 3.1 Chromatin immunoprecipitation experimental approach. The key experimental steps for ChIP (a-g). Native ChIP does not require the cross link step (a).

3.2 Results

3.2.1 Validating topoisomerase ChIP

3.2.1.1 Validating topoisomerase antibodies

The success of many molecular biology techniques is dependent on high quality antibodies specific to the protein of interest. This is particularly important for ChIP, where non-specific binding could identify false positive binding sites, high background or a poor signal to noise ratio. To ensure specificity seven topoisomerase antibodies were tested by western blot to confirm the detection of a single protein and immunofluorescence to ensure the nuclear localisation of this protein (Table 3.1).

Antibody	*Western blot dilution*	*Nuclear localisation*	*Supplier*
Anti - topo I (c21)	1:5000	Yes	Santa Cruz sc-32736
Anti - topo I	1:1000	No	Abcam ab3825
Anti - Scl70	1:5000	Yes	Patient serum
Anti - topo IIα (K19)	1:1000	Yes	Santa Cruz sc-5347
Anti - topo IIα	1:500	No	Cell Signalling #4733
Anti - topo IIβ	1:5000	Yes	BD Biosciences 611493
Anti - topo IIβ	1:100	No	Sant Cruz sc-5353
Anti - HP1α	1:5000	Yes	Chemicon MAB3584
Anti - GAPDH	1:5000	Yes	Cell Signalling 14C110
Anti - phospho-H2AX	1:1000	Yes	Upstate 05-636

Table 3.1 Antibodies tested by western blot and immunofluorescence. Western blot dilution for primary antibody incubation overnight at 4°C. Immunofluorescence yes/no indicates nuclear localisation (cytoplasmic in the case of GAPDH).

Both topoisomerase I antibodies tested identify a single protein at the correct molecular weight of 91 KDa by western blot (Figure 3.2a). However, the Santa Cruz antibody detected topoisomerase I with much higher affinity than the Abcam antibody. This Santa Cruz antibody has nuclear localisation by immunofluorescence (Figure 3.2a). Together this data supports the specificity of Santa Cruz anti-topoI (C21) for subsequent experiments.

A further antibody to topoisomerase I was isolated from the blood of a patient with the SCL70 form of scleroderma (Guldner et al., 1986). This antibody detects a band consistent with topoisomerase I alongside several other proteins (Figure 3.2d). The most prominent band is at 70 KDa, which is consistent with the proteasome degraded form of topoisomerase I produced following enzyme activity (Guldner et al., 1986; Tomicic and Kaina, 2013). Why scleroderma patients produce antibodies to a minority isoform of topoisomerase I remains unknown. Other bands, such as that ~50 KDa, probably detect additional auto-antibodies commonly observed in scleroderma such as those of centromeric proteins (Hamdouch et al., 2011). Together, this data suggests that the SCL70 antibody is not suitable for a specific topoisomerase I ChIP.

For topoisomerase IIα and IIβ a total of four antibodies were tested by western blot and immunofluorescence. The topoisomerase IIα from Cell Signalling and topoisomerase IIβ antibody from Santa Cruz did not identify any bands by western blot under our experimental conditions. The topoisomerase IIα antibody from Santa Cruz identifies a single band at ~170 KDa whilst the topoisomerase IIβ antibody from BD Biosciences identifies a single band at ~180 KDa (Figure 3.2b and 3.2c). Additionally, these antibodies detect proteins with nuclear localisation by immunofluoresence. This demonstrates the specificity of these topoisomerase II antibodies for use in subsequent experiments.

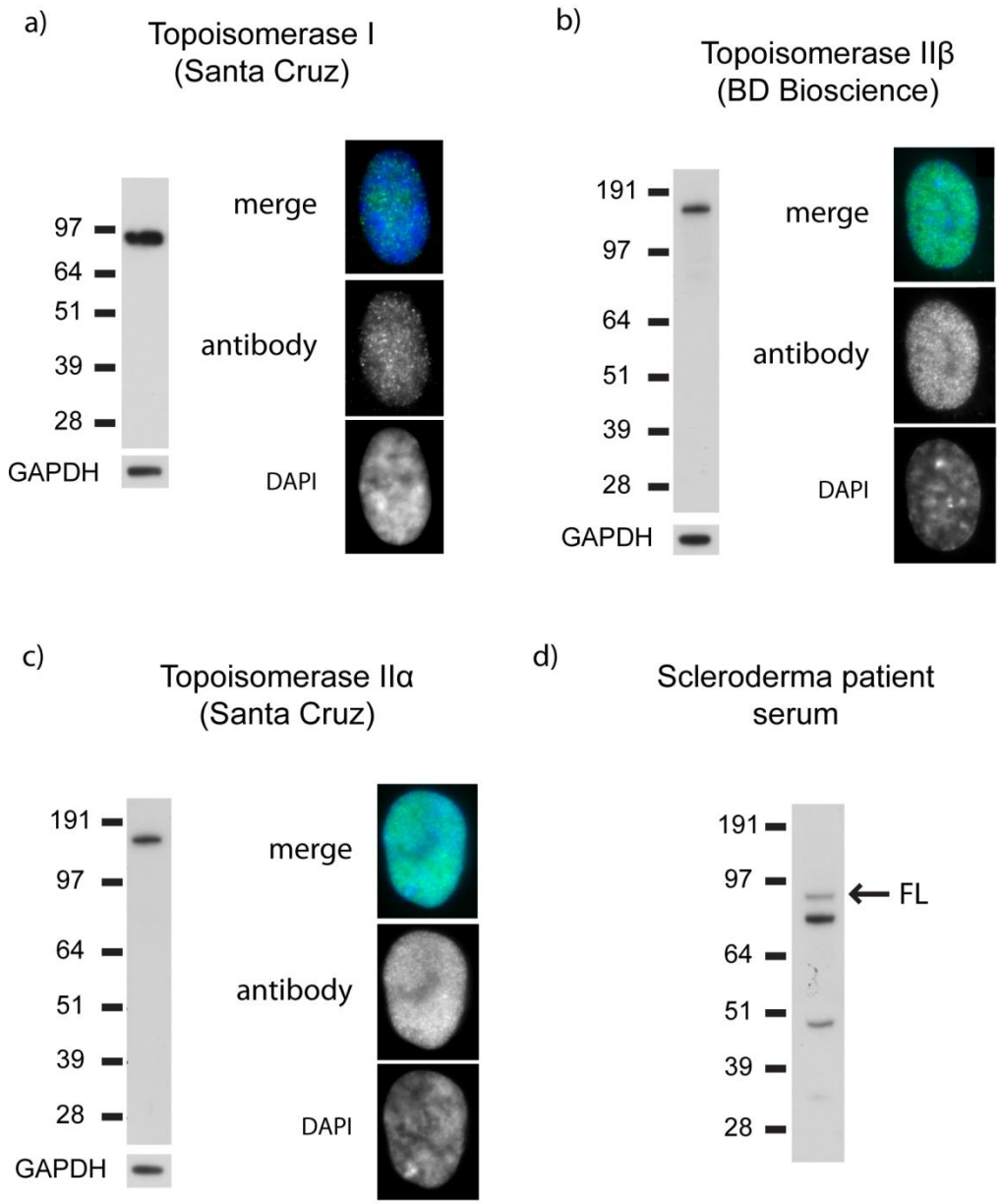


Figure 3.2 Topoisomerase antibody validation. a) - c) Western blot and immunofluorescence validation of antibody specificity. d) Scleroderma patient serum detects several isoforms/proteins. FL shows the full length topoisomerase I protein.

3.2.1.2 Validating topoisomerase chromatin interaction

ChIP experiments are reliant on stable interactions between the chromatin associated protein and DNA at specific loci. To identify what proportion of each topoisomerase is in a stable protein complex a salt extraction method was used (Henikoff et al., 2009). Proteins that form loose associations with DNA dissociate at lower NaCl concentrations than tightly bound proteins. In this experiment cell and nuclear membranes were lysed in a buffer containing 75, 150, 300, 500 or 1000 mM NaCl and large insoluble protein and nucleo-protein complexes were separated from soluble free protein and DNA by centrifugation. Western blot analysis identified that the majority of topoisomerase is not chromatin associated at all salt concentrations (Figure 3.3). At 150 mM NaCl the fraction of soluble topoisomerase is five fold higher for topoisomerase I, three-fold higher for topoisomerase IIα and two-fold higher for topoisomerase IIβ than the insoluble fraction. On the other hand, the tightly chromatin-associated HP1α protein is 2 fold higher in the insoluble fraction than in the soluble fraction. This identifies that the majority of topoisomerase in the cell is not stably associated in large protein or nucleo-protein complexes at the physiological salt concentration of nuclei (80 mM).

The minority of topoisomerase that forms insoluble complexes is likely to be critical for genome stability and regulation. One hypothesis is that topoisomerases form a structural component of the chromatin which, together with the nuclear matrix, regulate DNA structure within large DNA loops (Earnshaw and Heck, 1985). In general, proteins that maintain chromatin structure are tightly associated with the nucleo-protein complex. In the high salt fractions of the salt extraction assay (300 mM – 1000 mM) topoisomerases dissociate and become soluble (Figure 3.3), indicating that these proteins are less strongly associated in complexes than HP1α (Figure 3.3) or linker histones, which dissociate at 500 mM (Van Holde, 1989). Because topoisomerases are not tightly associated, this supports the use of chemical cross-linking to stabilise interactions in subsequent ChIP experiments.

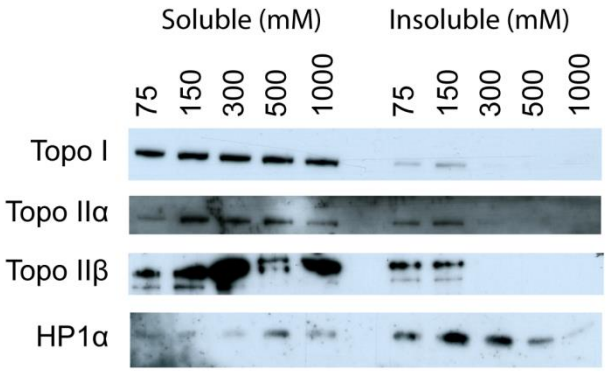


Figure 3.3 A minority of topoisomerases form stable interactions with chromatin under physiological salt conditions. Western blots for topoisomerase I, IIα and IIβ with HP1α as a positive control for protein association with chromatin. Soluble and insoluble fractions display the concentration of NaCl in mM.

To identify a level of cross-linking that stabilises topoisomerases in insoluble complexes, cells were incubated in different concentrations of formaldehyde and analysed by the salt extraction method. Cell lysates were prepared in the presence of 500 mM NaCl, which solubilises all topoisomerase under native conditions (Figure 3.3) and matches the most stringent immunoprecipitation wash condition. Pre-treating the cells with 0.5% or 1% formaldehyde prior to cell lysis stabilises all topoisomerase I in insoluble complexes (Figure 3.4). At lower formaldehyde concentrations topoisomerase I is present in both the soluble and insoluble fractions. In contrast, only a minority of the cytoplasmic protein GAPDH forms insoluble complexes at high formaldehyde concentrations (Figure 3.4). This suggests that the shift of topoisomerase I from generally soluble to generally insoluble following formaldehyde treatment is a property of the nucleus, consistent with an increase in chromatin association rather than general protein agglomeration.

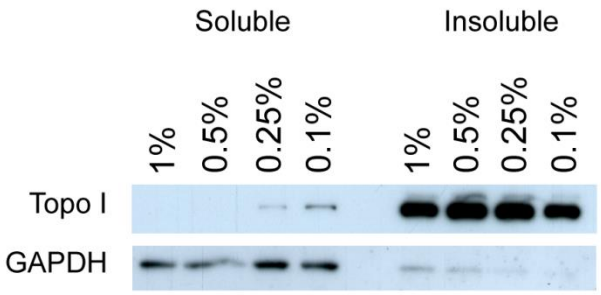


Figure 3.4 Formaldehyde cross-linking stabilises topoisomerase chromatin interactions. A western blot for topoisomerase I and GAPDH in non-chromatin (soluble) and chromatin (insoluble) fractions, from formaldehyde treated (%) cell lysates obtained in a buffer containing 500 mM NaCl.

The identification of topoisomerases in the insoluble fraction of the salt extraction assay identifies that they form protein and/or nucleo-protein complexes in the cell nucleus. To identify if topoisomerases are associated with chromatin, as would be expected from their function in the relief of DNA supercoiling and knots, sucrose gradient sedimentation was performed on isolated chromatin fragments. In this experiment, sucrose gradients separate protein, DNA and protein-DNA complexes based on their sedimentation rate, which is determined by mass and macro-molecular structure (Gilbert and Allan, 2001). Chromatin was isolated from nuclei by partial digestion with MNase followed by an overnight incubation in a hypotonic solution that pierces the nuclear membrane and permits the diffusion of chromatin fibres into the supernatant. It is important to optimise the MNase digestion for each cell type to obtain an optimum fragment length of ~ 1 kb. Based on a digestion time-course

(Figure 3.5a) the digestion was optimised to 10 minutes at 10 U/ml. Digested chromatin is sedimented on a 10-50% sucrose gradient at high speed and fractionated from low to high sucrose concentration. Chromatin complexes are found in fractions 2 through 5 as identified by the presence of DNA and protein. The spectrophotometer trace identified a peak of DNA in fractions 4 and 5 (Figure 3.5b), which is consistent with agarose gel electrophoresis on DNA purified from the fractions (Figure 3.5c). In addition, lower levels of DNA occur in fractions 2, 3 6 and 7 (Figure 3.5c). The distribution of protein, determined by coomassie stained SDS-PAGE, shows a similar distribution to that of DNA (Figure 3.5c). Fractions 6 and 7 contain DNA and histone proteins only, indicating that much of the protein complement of the chromatin has been stripped during the sedimentation process. Together, this data indicates that proteins associated with chromatin should be in fractions 2 through 5. To identify if topoisomerases are associated with the chromatin, western blots were performed on protein samples taken from the sucrose gradient fractions (Figure 3.5d). Consistent with chromatin association, topoisomerase I and IIβ are found in fractions 2 to 5. As positive controls HP1α and histone H3 western blots were performed and confirm the co-localisation with chromatin associated fractions. This data complements the salt extraction data, indicating that at least a proportion of the insoluble topoisomerase is chromatin associated, and supports subsequent experiments to determine the distribution of topoisomerases on chromatin by ChIP.

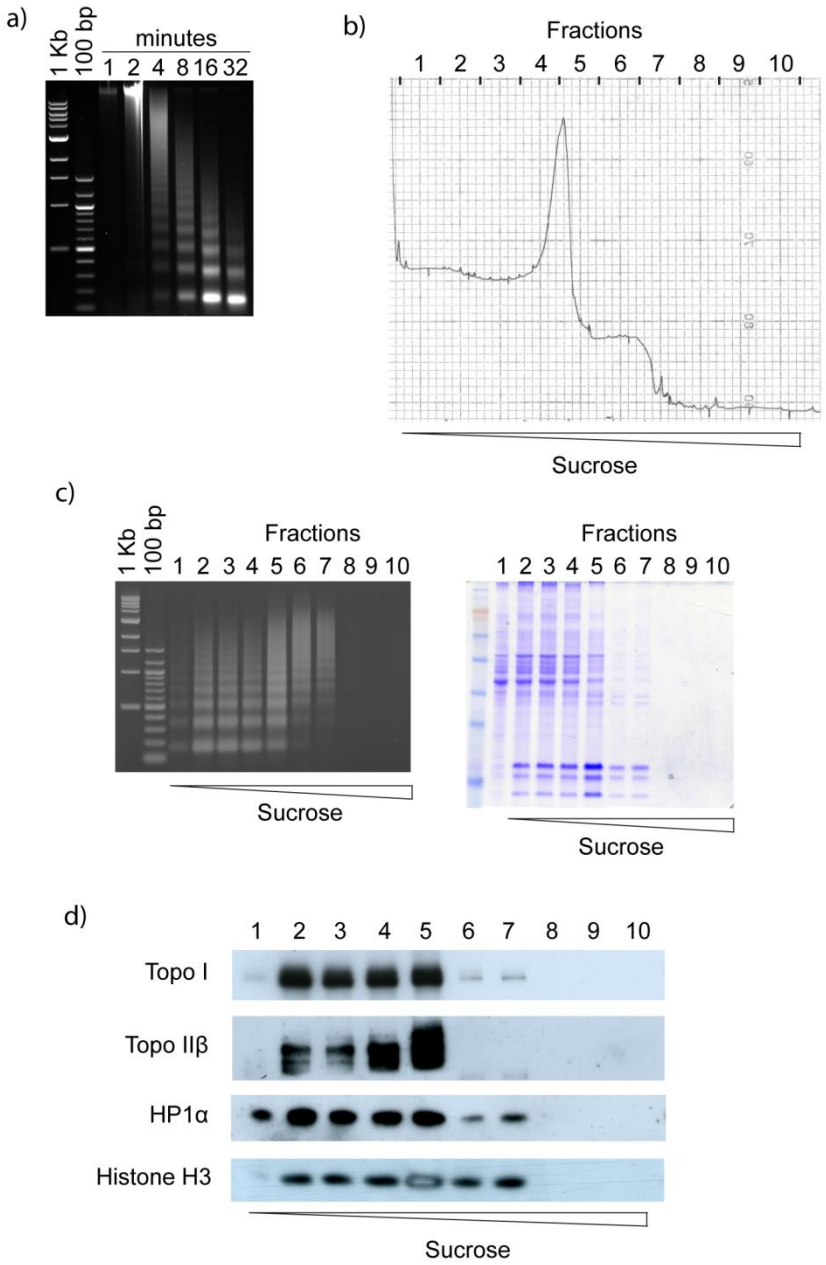


Figure 3.5 Topoisomerases are enriched in human chromatin. a) MNase digestion series for chromatin preparation. b) DNA trace measured by spectrophotometer at A260 during sucrose gradient fractionation. c) DNA content by agarose gel electrophoresis (left) and protein content by coomassie stained SDS PAGE for the sucrose gradient fractions. d) Western blot of sucrose gradient fractions for topoisomerase I, topoisomerase IIβ, HP1α and histone H3.

3.2.1.3 Optimising sonication conditions

The resolution of ChIP is depdendent upon DNA fragments size, with fragments of 250-500 bp giving the optimum compromise between chromatin integrity and mapping resolution. The efficiency of chromatin fragmentation is dependent on the complete lysis of the cell and nuclear membrane and the number of cycles of sonication used. To identify the optimum sonication conditions whole cell extracts and nuclear extracts were sonicated in a buffer containing 1% or 0.1% SDS for 6, 10 or 15 cycles of 30 seconds (Figure 3.6). The difference between whole cell and nuclear extract was minimal. However, the sociation buffer SDS concentration had a dramatic effect on DNA fragmentation efficiency, with 1% SDS showing much higher efficiency that 0.1% SDS. The most important factor for obtaining the optimum fragment size is the number of cycles of sonication, with 6 cycles giving an average fragment size of 1 kb, 10 cycles giving an average of 500 bp and 15 cycles giving an average of 300 bp in 1% SDS sonication buffer. Subsequent experiments identified that further sonication beyond 13 cycles had little effect on DNA fragment size (data not shown). Therefore, 13 rounds of sonication on whole cell extracts in a sonication buffer containing 1% SDS gave an optimum chromatin fragment size of 250-500 bp.

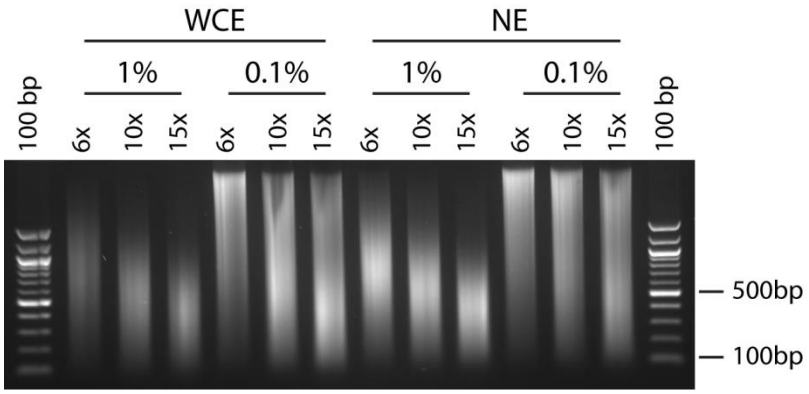


Figure 3.6 Optimising chromatin sonication conditions. 1.2% TBE agarose gel for sonicated whole cell extract (WCE) and nuclear extract (NE). The sonication buffers contained either 1% SDS (1%) or 0.1% SDS (0.1%). The extracts were sonicated for 6, 10 or 15 cycles (6x, 10x, 15x).

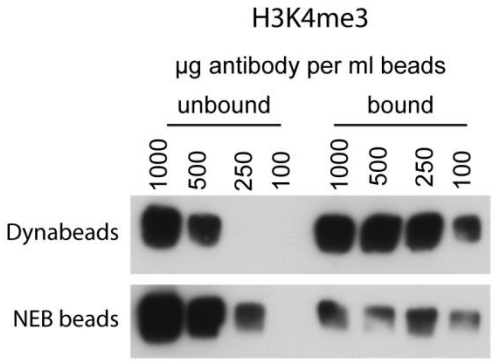
3.2.1.4 Optimising immunoprecipitation conditions

The choice of immunopreciptation reagents for the isolation of antibody-protein-DNA complexes can further affect the efficiency of a ChIP reaction. The most common protocols use agarose or paramagnetic beads coated in proteins that bind a specific immunoglobulin. For example, protein A derived from the bacterium *Staphylococcus aureus* binds most efficiently to the human immunoglobulins IgG1 and IgG2 (Johnstone, 1996). To test the efficiency of antibody-bead interactions immunogenic beads were incubated in solutions containing different antibody concentrations and the amount of bound antibody quantified by western blot. For example the protein A dynabeads bind more than half of the H3K4me3 antibody at 500 µg of antibody per 1 ml of beads giving a binding capacity of 300 µg/ml (Figure 3.7a). The NEB protein beads on the other hand bind around half of the antibody at 250 µg of antibody per 1 ml of beads giving a binding capacity of 125 µg/ml. Therefore, the dynabeads have more than twice the binding capacity of the NEB beads. This was also confirmed for topoisomerase antibodies binding to protein G

dynabeads (Figure 3.7b). This data suggests protein G dynabeads are optimal for the topoisomerase IIα and topoisomerase IIβ ChIP experiments.

The most suitable topoisomerase I antibody for my experiments was an IgM antibody, which has a different protein structure and has a high binding affinity for protein L coated beads. In developing the topoisomerase I approach protein L and anti-IgM beads were only available as agarose beads. The binding capacity of both anti-IgM and protein L agarose beads was higher than the paramagnetic beads, with 500 µg/ml and 800 µg/ml bound repsectively (Figure 3.7b). Therefore, for subsequent topoisomerase I ChIP experiments I used protein L coated agarose beads.

a)



b)

Antibody	*Bead type*	*Binding capacity (µg/ml)*
H3K4me3	Protein A Dynabead	300
	Protein A NEB beads	125
Topoisomerase IIα	Protein A Dynabead	200
	Protein A NEB beads	100
Topoisomerase IIβ	Protein A Dynabead	125
	Protein A NEB beads	0
Topoisomerase I	Protein L agarose	800
	Anti-IgM agarose	500

Figure 3.7 Antibody binding capacity of immunogenic beads. a) H3K4me3 binding capacity quantified by western blot. b) Western blot quantification identifies the binding capacity of antibody-bead combinations.

3.2.2.6 Topoisomerase inhibitors reduce topoisomerase protein concentration.

The topoisomerase drugs camptothecin and ICRF193 stabilise DNA-protein interactions by preventing the release of topoisomerase I or topoisomerase II respectively from the DNA. Camptothecin prevents the release of topoisomerase I by stabilising the bond formed between the enzyme and the 3' strand of the DNA, whereas ICRF193 prevents the conversion of topoisomerase II from the closed-clamp to the open-clamp form, in both cases trapping the enzyme on the DNA (Pommier, 2006; Roca et al., 1994). These drugs have both been shown to enhance the ChIP signal at sites where topoisomerases are active (e.g. Käs and Laemmli, 1992; Sano et al., 2008). It has been established that prolonged exposure to camptothecin and ICRF193 can result in protein loss by proteasomal degradation (Desai et al., 1997; Isik et al., 2003). A 50% reduction in topoisomerase I and IIβ was observed in protein extracts from cells treated with camptothecin and ICRF193 respectively (Figure 3.8). On the other hand, topoisomerase IIβ was unaffected by camptothecin treatment and topoisomerase IIα was unaffected by camptothecin or ICRF193 treatment. Interestingly, topoisomerase I levels increase in the presence of the topoisomerase II inhibitor ICRF193. Whether this upregulation performs a redundant functional role or is signalled by changes in DNA supercoiling following topoisomerase II inhibition is unknown.

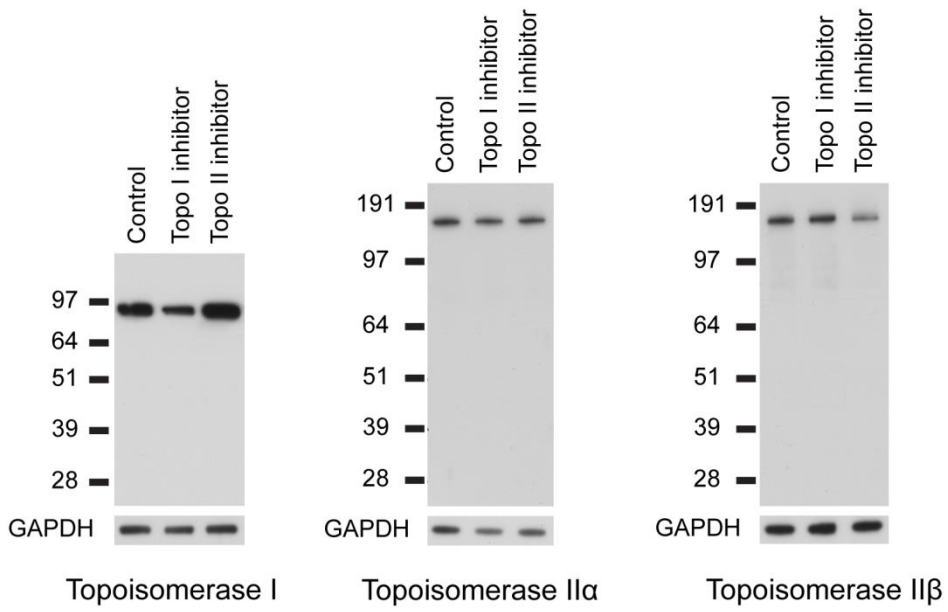


Figure 3.8 Topoisomerase I and IIβ protein reduced by topoisomerase inhibitors. Topoisomerase western blots in the presence of camptothecin (topo I inhibitor) or ICRF193 (topo II inhibitor). GAPDH shows even loading between samples.

3.2.3 Topoisomerase ChIP

Having optimised the conditions for topoisomerase I, IIα and IIβ ChIP a series of experiments were performed to investigate the distribution of topoisomerases across selected loci of the human genome. ChIP experiments were performed on untreated cells and cells treated with inhibitors that covalently associate topoisomerases with DNA. For each experiment ChIP enriched 'sample' and a control 'input' DNA was amplified, labelled with Cy3 or Cy5 fluorescent dyes and hybridised to custom microarrays tiling interesting regions of the human genome (Section 2.7). Microarrays were scanned and quality control analysis performed at the VUMC microarray facility, generating the data for subsequent bioinformatic analysis.

3.2.3.1 Raw data analysis

To confirm that microarray hybridisations were of high quality the scanned images were inspected for localised spatial defects in fluorophore signal intensity, which can be caused by manufacturing defects or user error. An analysis of both channels for each array confirmed the hybridisations were of high quality and appropriate for subsequent analysis.

The comparison of fluorophore intensity in two colour microarray analysis can be strongly influenced by the different spectral properties of Cy3 and Cy5, as well as their varied response to experimental and environmental variation. One striking example is the specific degradation of Cy5 by ozone, which can occur at levels found in the laboratory (Branham et al., 2007). To test the relative spectral properties of Cy3 and Cy5 in the topoisomerase ChIP microarrays, the fluorescence intensity distributions were plotted for each microarray (Figure 3.9a). The signal intensity distribution indicates a highly similar fluorescence distribution in the raw data within and between arrays (Figure 3.9a). However, the arrays still suffer from an intensity dependent bias universal in two colour microarray experiments, which is a result of the fluorescent properties of the dyes. A visualisation of this bias is presented in an MA plot, such as that for topoisomerase I sample A1 presented in Figure 3.9b. In this plot the log signal ratio (log2(R)/log2(G)) is plotted against the average signal of

the two fluorophores ((log(R)+log(G))/2). In an unbiased signal distribution the observed signal intensity ratio would not vary with the average signal intensity ratio and the plot would follow a horizontal line through zero. It is clear that the raw data does not follow this distribution for topoisomerase I sample A1 (Figure 3.9b), particularly at lower signal intensities, and a similar bias was present in each microarray. To reduce this technical bias several normalisation algorithms have been developed specifically for the analysis of microarray data.

Variance stabilising normalisation (VSN) is based on the observation that the variance of signal intensity increases with mean signal intensity. This is clear in the MA plot for unnormalised topoisomerase IA (Figure 3.9b), with a distinctly wedge shaped signal intensity ratio. The measured intensities are transformed so that the variance becomes independent of the mean and calibrated to account for inter-array variation. The MA plot of VSN normalised topoisomerase IA has a more even distribution over the full range of signal intensities (Figure 3.9b) Furthermore, the intensity dependent bias of the raw data is reduced dramatically, with the data being roughly distributed around an intensity ratio of zero across the range of average signal intensities. The relative distribution of signal intensity across arrays is also reduced, although several arrays no longer have the smooth normal distribution seen in the raw data (Figure 3.9a). An alternative two-step normalisation of the microarray data gave even better normalisation within and between microarrays, through the combined use of loess and scale normalisations. The loess normalisation adjusts for the intensity dependent bias observed in the MA plot by subtracting the bias estimated for a specified average signal intensity (i.e. adjusting the M value with respect to the A value). The loess normalisation corrects for technical bias between fluorophores within an array. To normalise between microarrays a scale normalisation is performed that standardises the distribution of all of the arrays by scaling the log-ratios to have the same median absolute deviation (MAD). To calculate the MAD the intensity of each probe is divided by the median deviation of all probes on the array. Following loess and scale normalisation there is no longer an intensity dependent bias in the MA plot (Figure 3.9b) and the relative distribution of signal intensity across arrays is more similar than in the unnormalised or VSN

normalised data (Figure 3.9a). Therefore, subsequent analysis was performed on loess/scale normalised data.

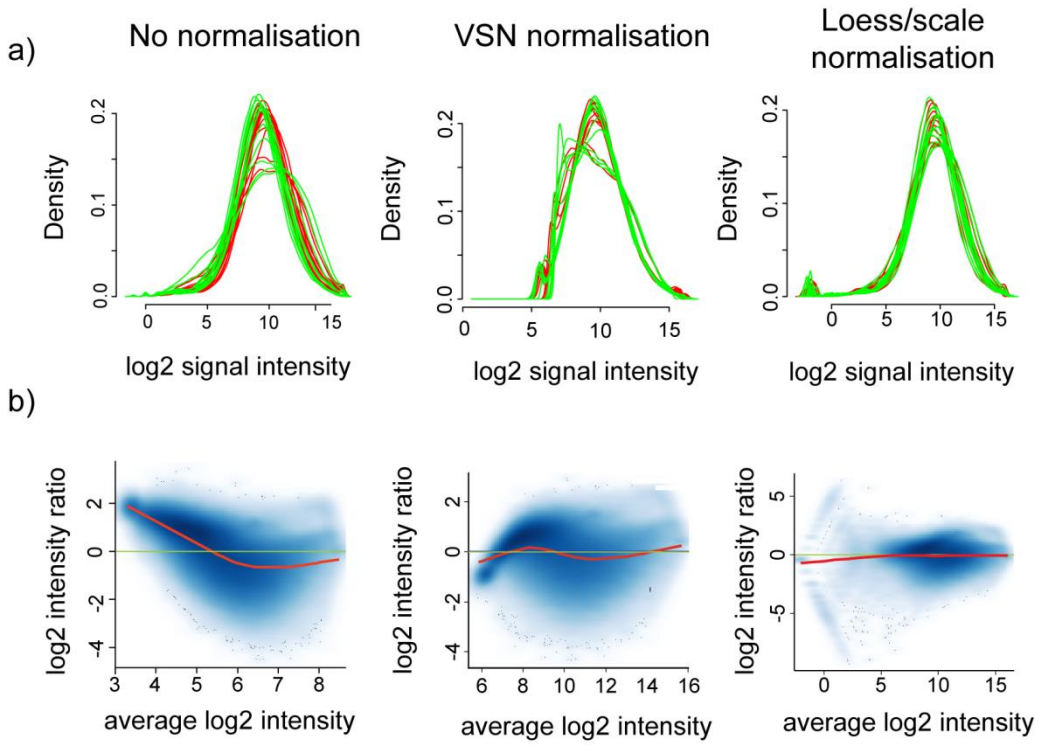


Figure 3.9 Microarray normalisation comparison. a) Signal intensity distribution of topoisomerase ChIP microarray Cy3 (green) and Cy5 (red) channels with and without normalisation. b) An example MA plot for topoisomerase I sample A1, indicating the intensity dependent bias prior to and following normalisation.

3.2.3.2 Inter-array variability

ChIP replicates show high similarity for each of the topoisomerase experiments, with topoisomerase I and topoisomerase IIβ having especially strong reproducibility of distribution between replicates (Figure 3.10). Treating the cells with camptothecin prior to topoisomerase I did not alter the distribution of topoisomerase I when compared to control samples (Figure 3.10). This indicates that camptothecin has not stalled topoisomerase at specific points within the loci studied. The camptothecin dependent degradation of topoisomerase I observed by western blot indicates that camptothecin is working in a manner similar to previous studies (Desai et al., 1997). Assuming the selection of gene rich/poor loci on the custom tiling arrays are representative of regions that topoisomerase I is likely to be active (see Section 2.7), there are a number of possible explanations for the similarity between camptothecin treated and non-treated samples. For example, the majority of topoisomerase I may already be in close spatial proximity with certain regions in the chromatin. This is supported by Dam-ID data in *Drosophila* where the distribution of topoisomerase I is identified in specific chromatin domains in an assay that identifies distribution independent of activity. Chromatin stability experiments suggested that the majority of topoisomerase I is not stably associated with the chromatin (Figure 3.3 and 3.5), despite its nuclear localisation (Figure 3.2). Together this data argues that topoisomerase I is enriched in certain regions of the genome and that camptothecin treatment does not enhance the detection of these regions. Therefore, because of the similarity between camptothecin and non-treated topoisomerase I ChIP distributions, these samples were treat as quadruplicates and the mean value of the four samples was used in subsequent analysis, represented as the mean distribution in Figure 3.10.

A similar situation is seen in the topoisomerase IIβ replicates, with a strong correspondence between replicates and between drug-treated and non-treated samples (Figure 3.10). The distribution of topoisomerase I and topoisomerase IIβ samples are clearly distinct, supporting a ChIP specific enrichment for each protein

in different regions of the genome. ICRF193 treated and non-treated samples were also treat as quadruplicates in subsequent analysis (mean distribution Figure 3.10).

The variability between topoisomerase IIα replicates was higher than observed for topoisomerase I and IIβ, giving a less clear domain distribution (Figure 3.10). Furthermore, one of the ICRF193 treated replicates failed at the hybridisation stage and was discounted from subsequent analysis. The mean distribution of the three remaining samples has a similar distribution to that of topoisomerase IIβ (Figure 3.10), consistent with the observations of Cowell et al. (2012). Subsequent analysis of topoisomerase IIα samples were based on the the mean distribution of these three samples.

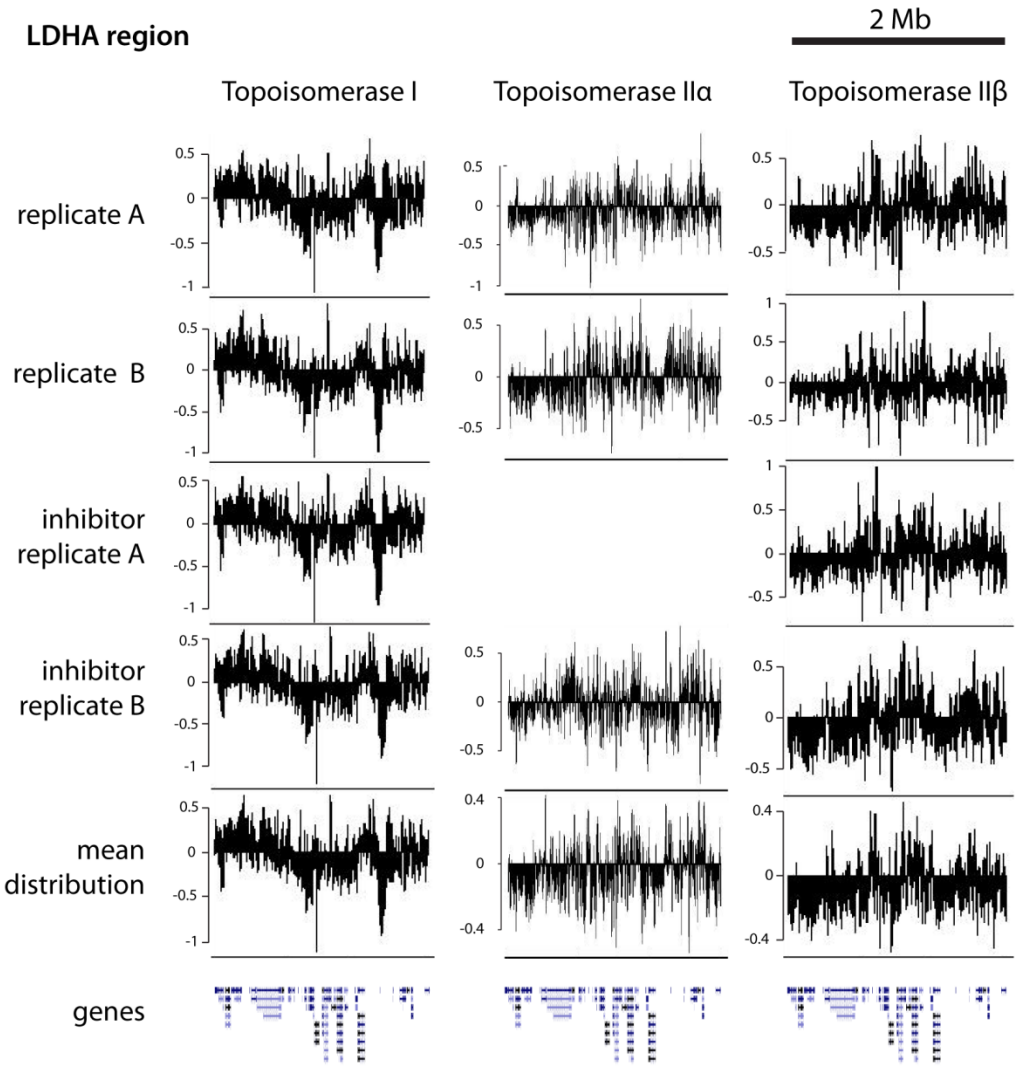


Figure 3.10 Topoisomerase ChIP microarray replicates show reproducibility. The distribution of topoisomerase ChIP data across the LDHA locus for each of the microarray experiments. Topoisomerase I and IIβ show high levels of reproducibility which is clearly reflected in the mean distribution. Topoisomerase IIα is more variable, but the mean distribution shows some similarity to the topoisomerase IIβ distribution.

3.2.4 Topoisomerases are enriched in large chromosomal domains

To identify the distribution of topoisomerase I, IIα and IIβ across loci the normalised probe intensities were submitted as custom tracks to the UCSC genome browser. This analysis identified domain scale enrichments of tens to hundreds of kilobases (Figure 3.11). The domains of topoisomerase I and II proteins appear to be distinct, whilst topoisomerase IIα and IIβ show similar distributions, as illustrated across the LDHA and IGBP1 loci. To determine the properties of these topoisomerase domains a rolling mean edge filter was used to determine the boundaries of topoisomerase enrichments (Materials and Methods Section 2.8.4.2). Stringent cut-offs were used to determine the points where a consistent enrichment becomes a consistent depletion, or vice versa, and the median value determined between boundaries. Topoisomerase domains were defined as regions with a positive median value, identifying 94 topoisomerase I, 112 topoisomerase IIα and 102 topoisomerase IIβ domains with a mean size of 86 kb, 95 kb and 102 kb respectively. These domains, displayed as coloured bars on Figure 3.11, were used for subsequent analyses of the inter-relationships between topoisomerases and the relationship between topoisomerase enzymes and other properties of DNA sequence, structure and function.

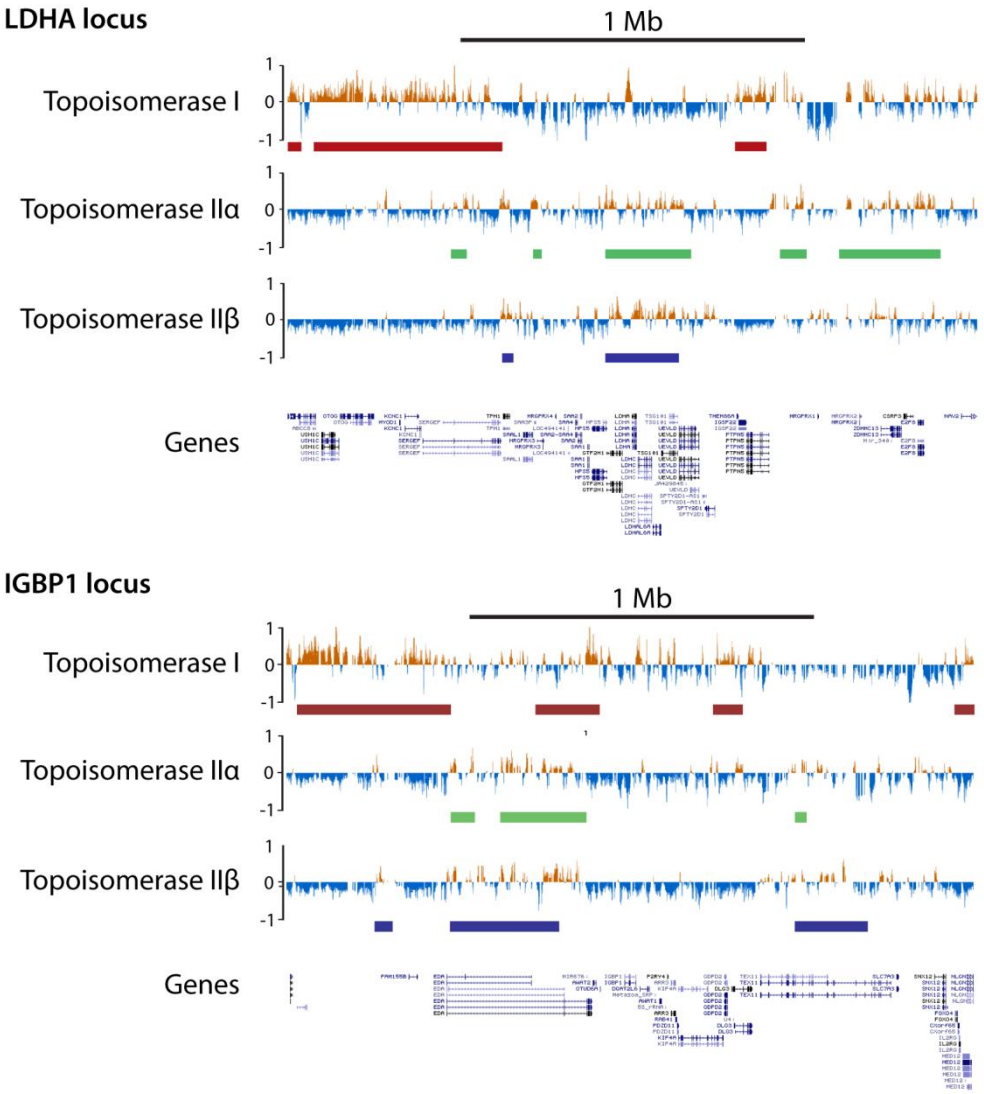


Figure 3.11 Topoisomerases are enriched at a domain scale. Topoisomerase I, IIα and IIβ ChIP enrichment across the LDHA and IGBP1 loci, with enriched domains marked as red, green or blue boxes respectively. Distribution taken from the UCSC genome browser using custom tracks smoothed with a 27 probe rolling median. Enriched domains calculated using an edge filter, as described in Materials and Methods Section 2.8.4.2. Y axis units – log2(sample/input) for each ChIP.

3.2.4.1 Properties of topoisomerase domains

To investigate the relative distributions of topoisomerase I, IIα and IIβ across human loci, the relative enrichment of each topoisomerase was identified for each type of topoisomerase domain (Figure 3.12). Topoisomerase IIα and IIβ domains show strong enrichment for one another, supporting observations in the UCSC genome browser that these proteins are enriched at similar positions (Figure 3.11). On the other hand, topoisomerase I is not strongly enriched at topoisomerase II domains, and topoisomerase IIβ is depleted with high significance at topoisomerase I domains. This indicates a separation of topoisomerase I and II across the genome and suggests that the activity of topoisomerase IIβ is not generally required at topoisomerase I enriched regions.

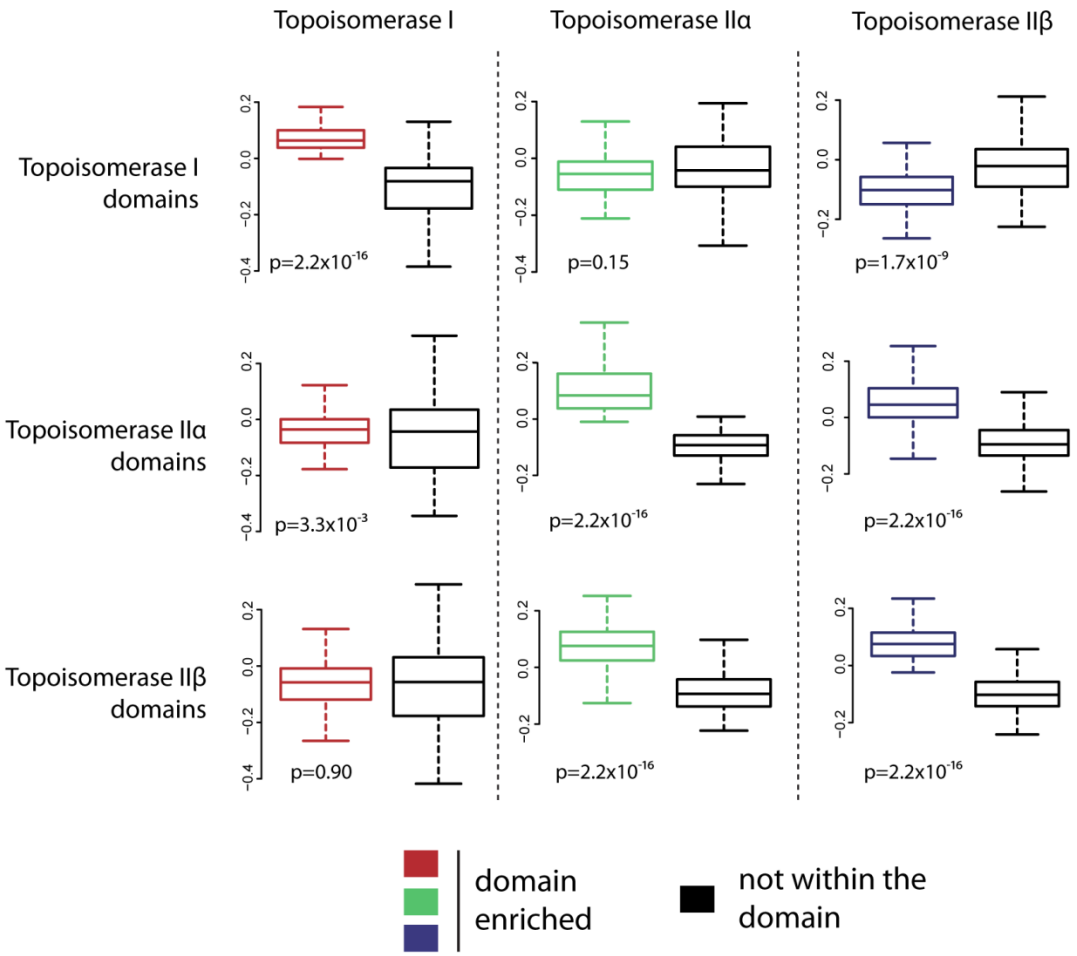


Figure 3.12 Topoisomerase enrichment within topoisomerase domains. Enrichment of topoisomerase I, IIα and IIβ in the topoisomerase domains. Black boxplots represent the probes not contained within the domain for each topoisomerase separately. Scale is log2 enrichment, p is a two-tailed students t test p-value.

To identify the sequence, structural and functional characteristics of the topoisomerase domains, which may account for the separate distributions of topoisomerase I and II enzymes, a comparison was made with online and in-house datasets. Comparing the nucleotide composition of topoisomerase domains identified that topoisomerase I domains are strongly enriched for GC, whereas topoisomerase IIα and IIβ domains are associated with more AT rich regions of the genome (Figure 3.13a). At a simple level this indicates that topoisomerase I domains are associated with gene rich regions, which are typically GC rich, whereas topoisomerase II domains are likely to be in gene poor regions. To test this directly the number of transcription start sites (TSSs) per 10 kb was quantified for each topoisomerase domain. Topoisomerase I domains are strongly enriched for TSSs when compared to topoisomerase II domains (Figure 3.13b), further supporting the relationship between topoisomerase I, high GC and genes.

The indication thus far is that topoisomerase I is associated with genes and topoisomerase II is associated with gene poor regions. To identify if this corresponds to a relationship between topoisomerase I and transcription, the relative enrichment of RNA polymerase II was identified for each type of topoisomerase domain using RNA polymerase II ChIP data from within the lab (Naughton et al., 2013a). RNA polymerase II was strongly enriched in topoisomerase I domains compared to topoisomerase II domains, with a weak significant enrichment at topoisomerase IIα compared to topoisomerase IIβ domains (Figure 3.13c). This argues that topoisomerase I is important for remodelling DNA structure at transcriptionally active regions within the genome (investigated further in Section 3.2.5).

The remodelling of DNA structure by topoisomerases is essential for efficient transcription and cell cycle progression, but the relationship between DNA supercoiling and topoisomerase distribution in humans is unknown. To identify this relationship the relative enrichment of psoralen in each of the topoisomerase domains was identified using psoralen-IP data generated in the lab (Naughton et al., 2013a) . Psoralen binds preferentially to under-wound DNA, giving a relative indication of DNA supercoiling across the genome (see Sections 1.2.5 and 4.1). Psoralen was

strongly enriched in topoisomerase I domains and depleted in topoisomerase IIα and IIβ domains (Figure 3.13d). This identifies that topoisomerase I domains have a more under-wound DNA structure, which is associated with transcription and a more accessible chromatin structure (Bermúdez et al., 2010; Matsumoto and Hirose, 2004; Naughton et al., 2013a), whilst topoisomerase II have an over-wound DNA structure associated with gene repression (Bermúdez et al., 2010; Matsumoto and Hirose, 2004; Naughton et al., 2013a).

Together this data identifies for the first time how topoisomerases are distributed within the human genome. The domain scale enrichment of topoisomerases is related to sequence, RNA polymerase II distribution and DNA structure. To further characterise the relationship between topoisomerases, RNA polymerase II and DNA structure I undertook a more detailed analysis.

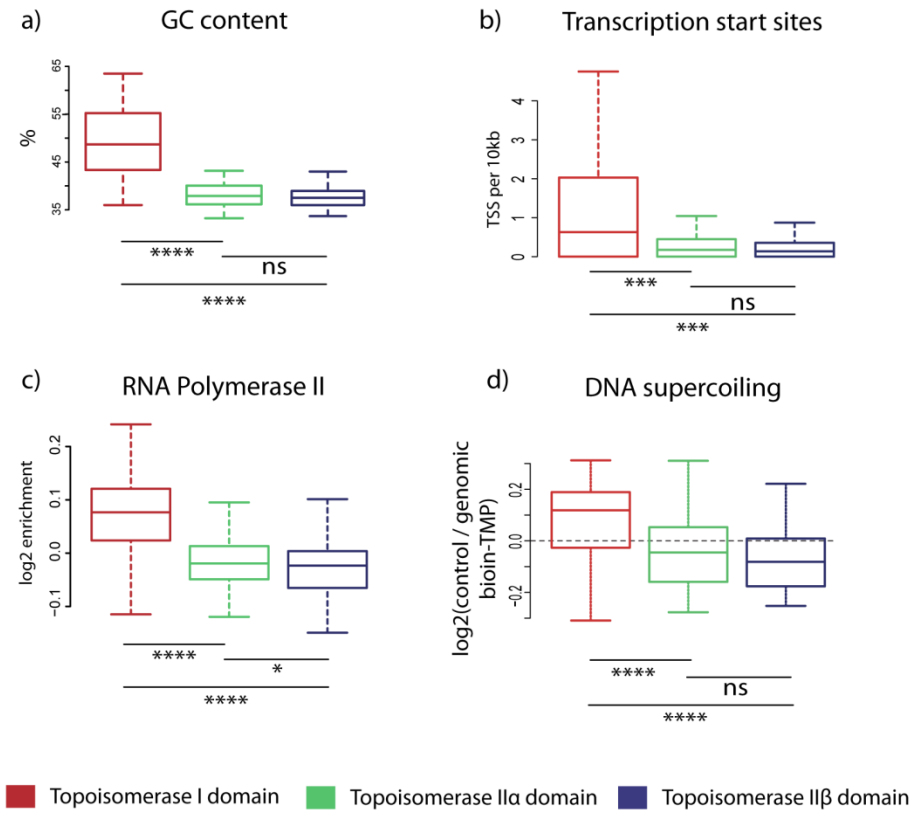


Figure 3.13 Sequence, functional and structural properties of topoisomerase domains. Topoisomerase I (red), IIα (green) and IIβ (blue) domain boxplots for a) GC percentage distribution, b) number of transcription starts sites per domain, c) RNA polymerase II enrichment and d) DNA supercoiling measured by psoralen enrichment. GC percentage distribution is measured per domain to give the overall distribution across domains. RNA polymerase II enrichment was derived from ChIP experiments and psoralen enrichment from psoralen-IP experiments performed in the lab for (Naughton et al., 2013a). The dotted line in d) represents the boundary for under-wound DNA (enriched for psoralen relative to genomic therefore above the line) and over-wound DNA (depleted for psoralen relative to genomic therefore below the line). Significance was identified by students t test with the p value represented by: 'ns' = not significant, '*'<0.05, '**'<0.005, '***'<0.0005, '****'<0.00005.

3.2.5 RNA polymerase II and topoisomerase I co-localise *in vivo*

There is a clear positive correlation between topoisomerase I and RNA polymerase II in the analysis of topoisomerase domains (Figure 3.13c). To further characterise this relationship, the distribution of topoisomerases and RNA polymerase was examined at a probe-by-probe scale. The distribution of RNA polymerase II data across the loci displays a strong relationship with topoisomerase I, at a domain scale and at the scale of individual genes (Figure 3.14a). A scatter plot confirms this close relationship, with the Pearson's moment correlation coefficient between topoisomerase I and initiating RNA polymerase II of 0.79. Therefore, the correlation between topoisomerase I and RNA polymerase II is comparable to the correlations between total and initiating RNA polymerase (Pearson's 0.78) (Figure 3.14b), which also has a similar distribution at a gene and domain scale (Figure 3.14a). Consistent with the domain scale observations, topoisomerase IIα shows a poor correlation with RNA polymerase II (Pearson's 0.26) and topoisomerase IIβ shows no correlation (Pearson's 0.09). This is reflected in the domain and promoter scale distributions (Figure 3.14a) and in the scatter plots (Figure 3.14b). Together this data identifies at high resolution the co-localisation of topoisomerase I and RNA polymerase II across megabase-scale loci of the human genome, and identifies no such general relationship exists for topoisomerase IIα or IIβ.

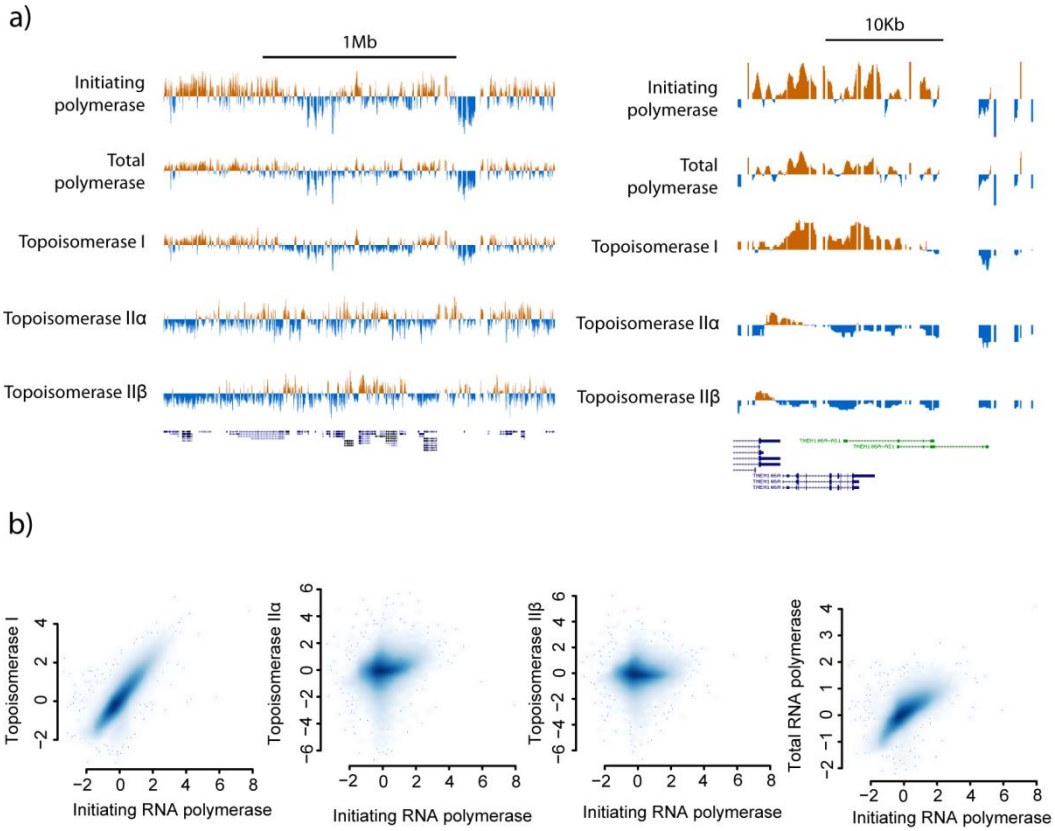


Figure 3.14 Relative distribution of topoisomerases and RNA polymerase II. a) Topoisomerase and RNA polymerase II distribution across the LDHA locus (left) and at the TMEM106 promoter (right). b) Scatter plots of the relationship between topoisomerase and RNA polymerase II distribution from ChIP-chip experiments. Initiating and total polymerase ChIP experiments performed by Naughton et al. (2013a).

3.2.6 Topoisomerase I and II are strongly enriched in distinct DNA supercoiling domains.

The observation that topoisomerase I domains are strongly enriched for under-wound DNA whereas topoisomerase IIα and IIβ domains are strongly enriched for over-wound DNA identifies a relationship between topoisomerase and DNA supercoil distribution (Figure 3.13d). To investigate the *in vivo* relationship between DNA supercoiling and topoisomerases further, the relative enrichment of topoisomerase I, IIα and IIβ was identified for the DNA supercoil domains characterised in Naughton et al. (2013). These domains classify the genome into regions that are under-wound, stable or over-wound using a psoralen-IP approach. Under-wound domains are associated with transcriptionally active GC rich sequences whereas over-wound domains are associated with inactive AT rich regions and stable domains have intermediate properties (Naughton et al., 2013a). Topoisomerase I is significantly enriched in under-wound DNA supercoil domains and depleted in both the stable and under-wound domains (Figure 3.15a and 3.15b), supporting the model that the major function of topoisomerase I is in the maintenance of DNA supercoils at transcriptionally active regions. Topoisomerase IIα and IIβ are both strongly enriched in over-wound DNA supercoil domains and depleted in under-wound domains (Figure 3.15a and 3.15b). Therefore, distinct DNA supercoil domains are serviced by either topoisomerase I or topoisomerase II proteins *in vivo*.

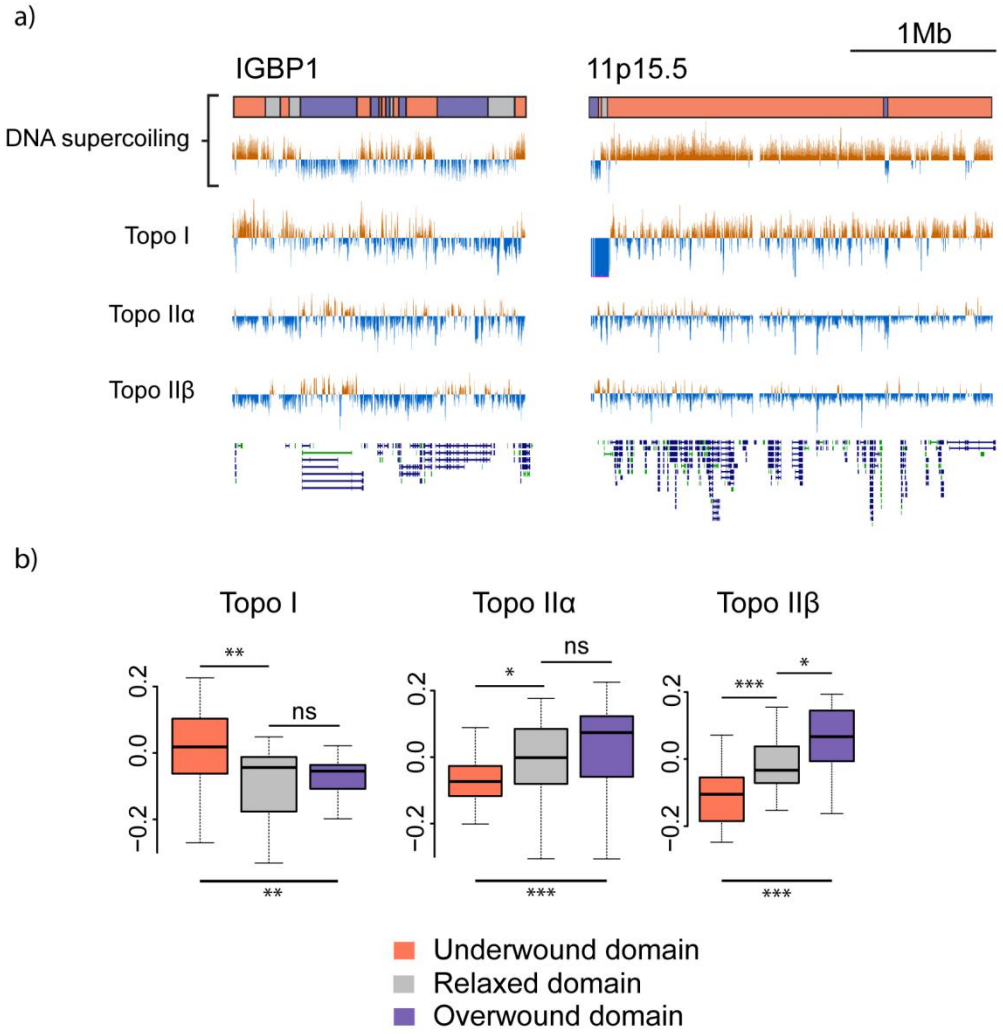


Figure 3.15 Topoisomerase enrichment in DNA supercoiling domains. a) DNA supercoil and topoisomerase distribution at the IGBP1 and 11p15.5 loci. DNA supercoiling analysis on data from C. Naughton. DNA supercoil domains data from Naughton et al. (2013). b) Boxplot of topoisomerase enrichemnt for the DNA supercoil domains. T-test p values are: 'ns' not singificant, '*' <0.05, '**' <0.005, '***' <0.0005.

3.2.7 Topoisomerase depletion at telomeres and common fragile sites

An analysis of the pattern of topoisomerase binding across loci identifies a number of sites at which the general relationships between topoisomerases and DNA supercoiling identified in Section 3.2.4 break down. At the telomere of chromosome 11 and fragile sites on chromosomes 1, 3, 16 and 17 there is an unusually strong depletion for topoisomerase I. In the case of the p-arm telomeric end of chromosome 11, there is a pronounced ~200 kb depletion of topoisomerase I followed by a low level enrichment 1 Mb downstream (Figure 3.16a). Unusually, this topoisomerase I enrichment corresponds to a domain of over-wound DNA and not with an enrichment for topoisomerase IIα or IIβ. The boundary of topoisomerase I enrichment and depletion does not correspond to a change in DNA supercoiling, indicating that the strong depletion for topoisomerase I does not significantly alter DNA structure at the chromosome end. Together this data identifies that at human telomeres, the general patterns of topoisomerase and DNA structure identified elsewhere in the genome do not apply. In yeast, the introduction of over-wound DNA supercoils does not affect gene expression in the region < 100 kb from the telomeric end and it is likely that this is due to the free diffusion of supercoils from the DNA (Joshi et al., 2010). This is consistent with the region of topoisomerase I depletion in human cells and indicates that human telomeres may also release DNA supercoiling independent of topoisomerase activity.

Another breakdown in the relationship between topoisomerases and DNA supercoiling is observed at the common fragile sites FRA3B and FRA16D (Figure 3.16b). In both CFSs there is a strong depletion for both topoisomerase I and topoisomerase II over a ~1 Mb region. Furthemore, at the RNU1 and RNU2 fragile sites, which are induced following adenovirus infection, a similar depletion is observed (data not shown). The coordinated depletion of topoisomerase I, IIα and IIβ is not observed for any other loci investigated. Furthermore, these sites have an unusual DNA structure, with a consistent over-wound DNA structure over several megabases. Together this data indicates that a lack of topoisomerase I, IIα and IIβ

enzymes could contribute to the genome instability observed at fragile sites. This hypothesis is investigated more thoroughly in chapter 5.

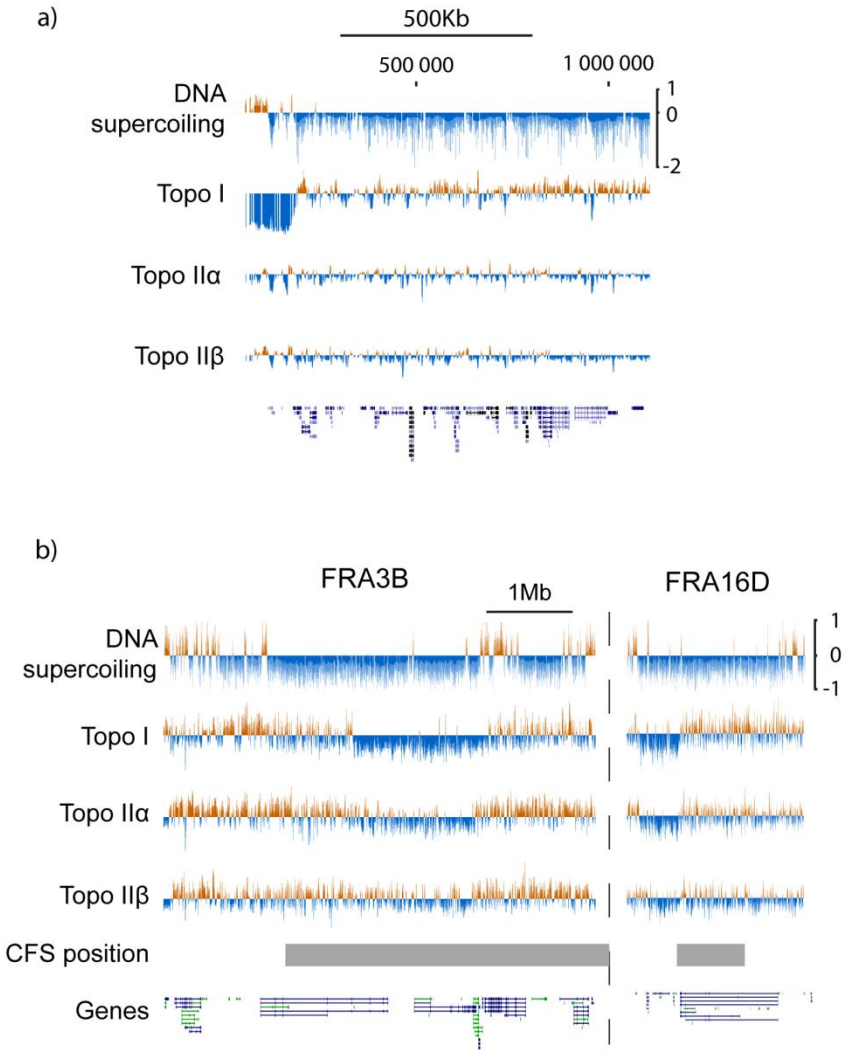


Figure 3.16 Topoisomerases are depleted at telomeres and common fragile sites. a) Topoisomerase I is depleted at the telomeric end of chromosome 11p. The plot shows the first 1.2 Mb of chromosome 11, with a scale bar to indicate genomic distance. The scale of enrichment/depletion is consistent between the four samples (-2 to 1 log2 fold enrichment). b) Topoisomerase I, IIα and IIβ are depleted at common fragile sites. The plots show the FRA3B and FRA16D loci to the same scale. The scale of enrichment is consistent between the four samples (-1 to 1 log2 fold enrichment). DNA supercoiling data generated by Naughton et al (2013). In both a) and b) DNA supercoiling is log2(control / genomic) bTMP.

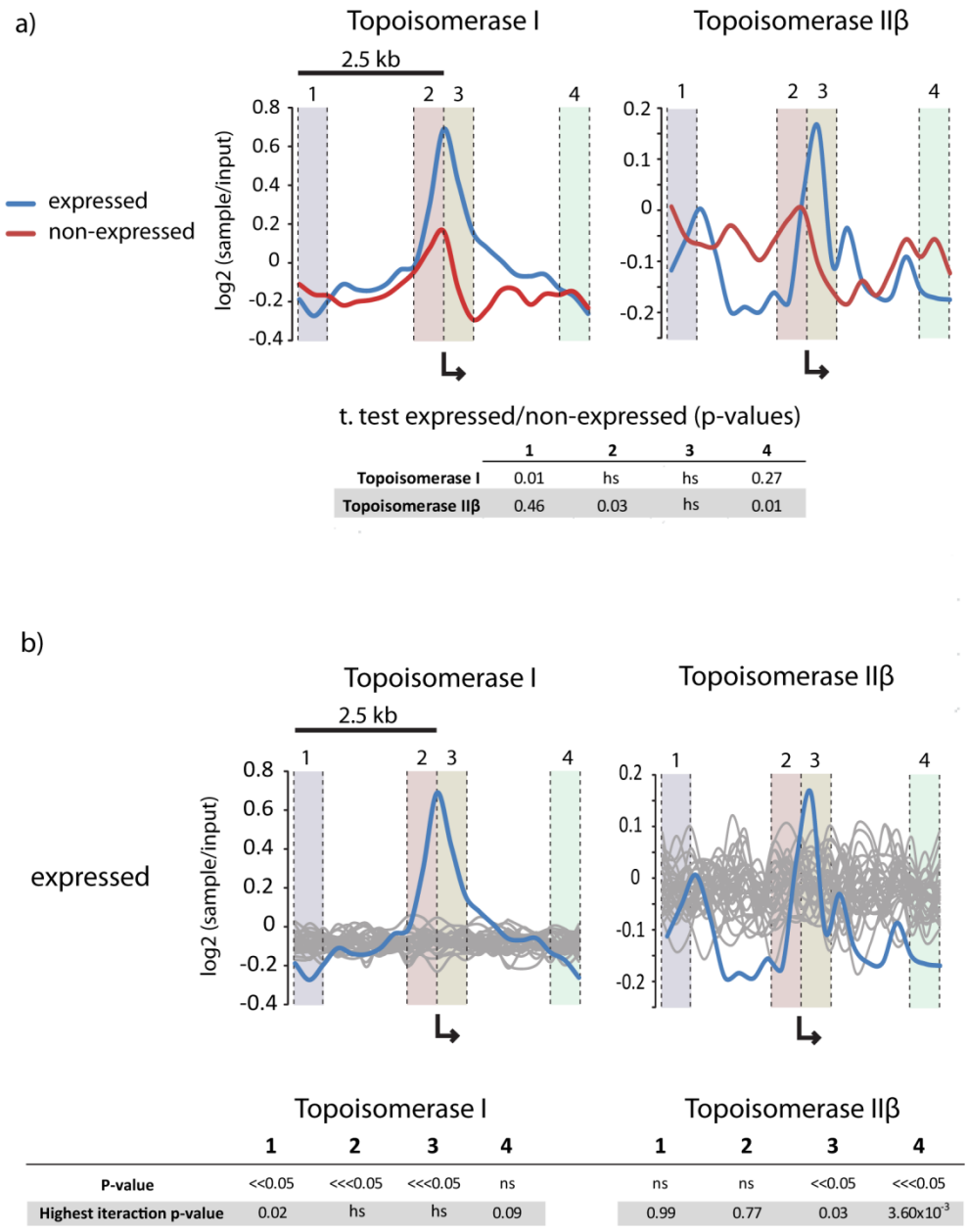
3.2.8 Topoisomerase distribution at promoters supports distinct biological function

The majority of coding and non-coding transcription occurs at gene promoter regions making promoters a hotspot for the generation of DNA supercoiling. The distribution of topoisomerases at promoter regions has been proposed to regulate and maintain DNA supercoiling at these sites, but the general distribution of topoisomerase I and II at human promoters remains unknown. Several studies have identified enhanced topoisomerase binding at transcription start sites (e.g. Cowell et al., 2012; Kouzine et al., 2013; Lyu et al., 2006; Sano et al., 2008) but no consensus as to the distribution of topoisomerases at TSSs has been established. Most recently Kouzine et al. (2013) proposed a model whereby topoisomerase I has a diffuse activity upstream of the transcription start site whereas topoisomerase II has a more focal distribution at the transcription start site, particularly in highly expressed genes. To test this model and establish the distribution of topoisomerase enzymes with respect to gene promoters the topoisomerase ChIP data was analysed with respect to gene promoters.

3.2.8.1 Topoisomerases have a focused enrichment at the transcription start site of expressed genes

Topoisomerase distribution at transcription start sites on the Agilent custom tiling arrays identified peaks of topoisomerase I, IIα and IIβ at the TSS relative to a position 2.5 kb upstream (data not shown), but further analysis of this data was limited by the small number of transcription start sites (372 TSSs) on the array. To investigate topoisomerase distribution at gene promoters in more detail, particularly with respect to gene expression, topoisomerase I and topoisomerase IIβ samples were hybridised to Nimblegen whole chromosome 11 microarrays (2,509 TSSs). These samples were chosen due to their clear reproducibility and distinct distributions in earlier analysis (section 3.2.3.2). Microarrays were analysed and normalised as described in Section 3.2.3.1.

To idenfity the distribution of topoisomerases at gene promoters, the mean ChIP enrichment over a 5 kb region centred on the transcription start site was identified for expressed genes (487 TSSs) compared to non-expressed genes (1161 TSSs). Topoisomerase I is clearly enriched at the transcription start site of expressed genes when compared with non-expressed genes (Figure 3.17a). Topoisomerase IIβ is enriched 500 bp downstream of the transcription start site in the expressed compared to the non-expressed genes. To identify if these enrichments could have occurred by chance, the distribution of topoisomerases around randomly selected points on the chromosome was performed for 487 points over thirty iterations. The difference in topoisomerase binding was identified for four ranges within the dataset by students t-test (-2500 to -2000, -500 to 0, 0 to 500 and 2000 to 2500 bp) with each iteration. After thirty iterations the topoisomerase I sample enrichment remained much more highly significant than the random samples at the transcription start site (Figure 3.17b). Furthermore, the upstream distribution of topoisomerase I was signficantly depleted compared to random iterations, making the TSS enrichment even more striking. The topoisomerase I enrichment extends a short distance into the body of expressed genes, but by 2.5 kb downstream it is not significantly different from a random distribution. Topoisomerase IIβ immediately downstream of the TSS (0-500 bp) is also significantly enriched compared to random iterations, although the magnitude of this difference is clearly smaller than for topoisomerase I (Figure 3.17b). Downstream of the TSS at 2-2.5 kb there is a significant depletion of topoisomerase IIβ, consistent with the general depletion of topoisomerase IIβ at gene rich regions (Section 3.2.4). This can be seen to continue upstream of the TSS, but at 2-2.5 kb there is a non-significant difference between topoisomerase IIβ and random iterations for expressed genes. What the significance of this upstream enrichment is not immediately obvious. Together this data identifies a clear enrichment of topoisomerase I and IIβ at the transcription start sites of expressed genes. The enrichment of both topoisomerases is focused at the TSS supporting the 'focal mode' hypothesis suggested in Kouzine et al. (2013) at active gene promoters.



	1	2	3	4
Topoisomerase I	0.01	hs	hs	0.27
Topoisomerase IIβ	0.46	0.03	hs	0.01

	Topoisomerase I				Topoisomerase IIβ			
	1	2	3	4	1	2	3	4
P-value	<<0.05	<<<0.05	<<<0.05	ns	ns	ns	<<0.05	<<<0.05
Highest iteraction p-value	0.02	hs	hs	0.09	0.99	0.77	0.03	3.60×10^{-3}

Figure 3.17 Focal enrichment of topoisomerases at the TSS of gene promoters. a) The distribution of expressed (blue) and non-expressed (red) genes over 5 kb centred on the TSS. To assess the significance between distributions, Student's t-tests were performed on four data ranges: -2500 to -2000, -500 to 0, 0 to 500 and 2000 to 2500. T-test values are recorded in the table for each of the data ranges. b) Peak significance determined by iteration analysis on random positions. To identify if the observed peaks

could have occurred by chance, the distribution of topoisomerases was identified around 30 sets of random probes and a t-test performed between sample and each set of random probes for the same four data ranges. The table records the highest p-value observed in the iterations, giving an indication of sample enrichment over random positions. P-values were calculated from the 30 iterations: ns – not significant, hs – highly significant ($p<1x10^{-10}$), $p<0.05$ – one p-value more than 0.05 out of 30, $p<<0.05$ – one or more p-values over 0.01 but none over 0.05, $p<<<0.05$ – no values over 0.01.

3.2.8.2 RNA polymerase II and under-wound DNA supercoiling are enriched with topoisomerase at expressed gene promoters.

To establish if the relationships between topoisomerase enrichment, gene expression, RNA polymerase II and DNA supercoiling that were observed for topoisomerase domains also hold true for gene promoters, the relative enrichment of RNA polymerase II and DNA supercoiling were analysed with respect to TSSs. Unsurprisingly RNA polymerase II is enriched at the TSSs of expressed genes compared to the distribution at non-expressed genes (Figure 3.18). This enrichment follows the same distribution as that observed for topoisomerase I, supporting the domain analysis (Section 3.2.4) and probe-by-probe similarity (Section 3.2.5). Furthermore, the small peak of enrichment upstream of the TSS that is present in both non-expressed topoisomerase I and RNA polymerase II is indicative of paused polymerase, in agreement with previous data (e.g. Kwak et al., 2013). On the other hand, there is no similarity between RNA polymerase II and topoisomerase IIβ distribution. Therefore, RNA polymerase II and topoisomerase I co-localise at

transcription start sites as well as on a larger-scale, supporting *in vitro* data that they work together during transcription to relieve DNA supercoiling.

To establish how DNA supercoil distribution is affected by RNA polymerase II and topoisomerase enrichment at TSSs, the relative enrichment of psoralen was identified for expressed and non-expressed genes. Both expressed and non-expressed genes show a relative under-winding across the 5 kb centred on the TSS (control minus genomic value > 0), with the greatest enrichment for expressed genes (Figure 3.18). The under-wound structure of expressed promoters is interrupted by a more over-wound structure at the TSS, an observation that is discussed further in Chapter 4. Together this data shows that there is a distinct expressed-promoter under-wound DNA stucture, in spite of the enriched topoisomerase distribution observed at these promoters.

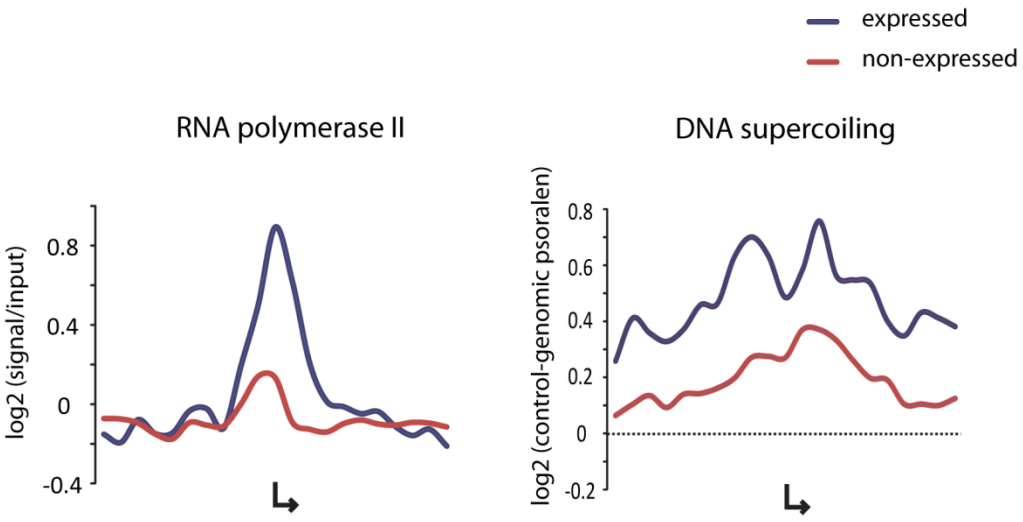
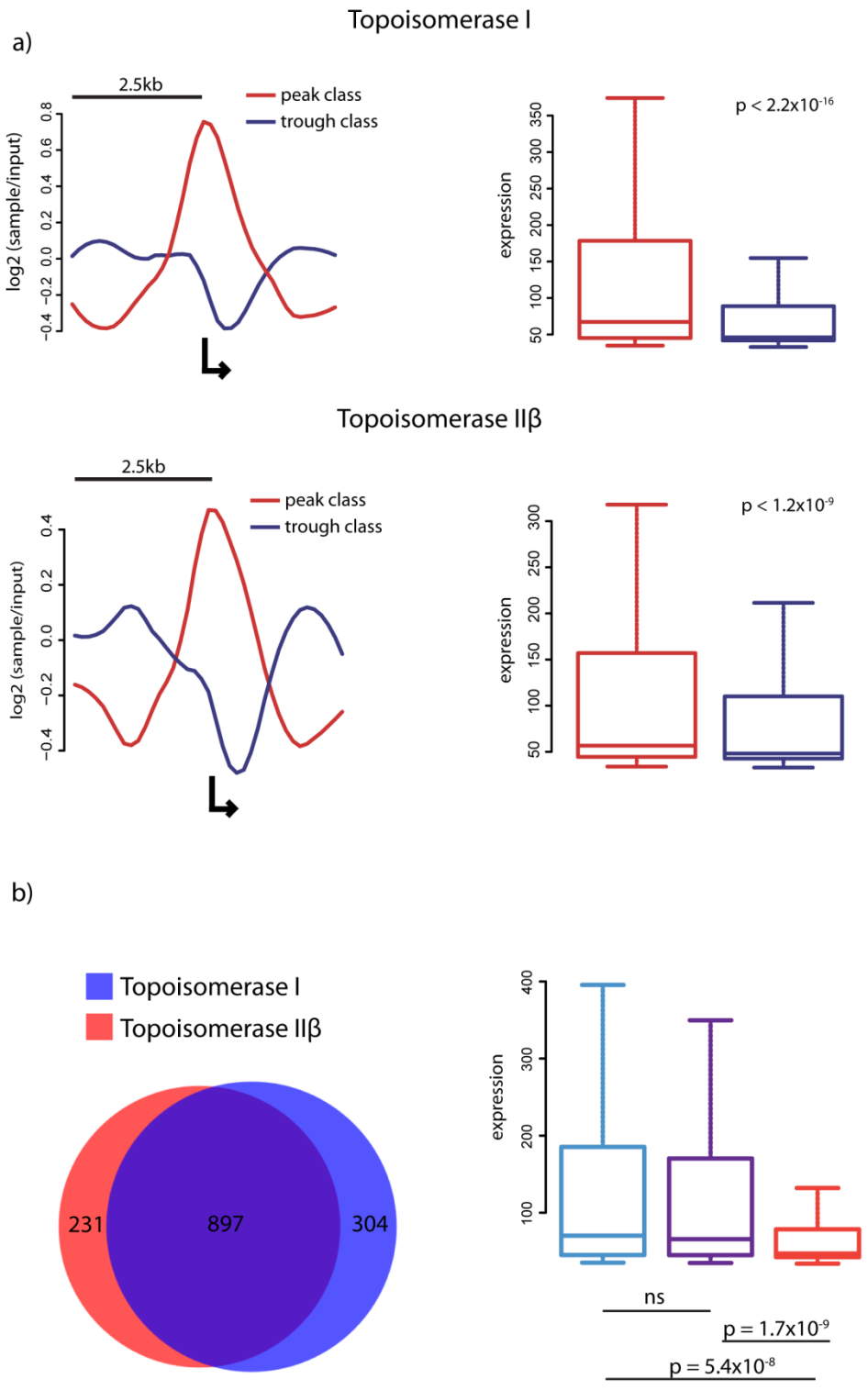


Figure 3.18 RNA polymerase II and under-wound DNA supercoils are enriched at TSSs in expressed genes. The distribution of RNA polymerase II and DNA supercoiling over 5 kb centred on the TSS. RNA polymerase expressed (blue) and non-expressed (red) shows a focal distribution around TSSs. Psoralen is enriched in chromatin compared to genomic DNA at both expressed (blue) and non-expressed (red) TSSs. DNA structure is more under-wound above the dotted line through zero and more over-wound below the line.

3.2.8.3 Topoisomerase IIβ peaks at promoters do not correlate with increased expression

To separate the role of topoisomerase I and II, promoters were identified with a peak of topoisomerase I only, topoisomerase IIβ only and both topoisomerase I and topoisomerase IIβ. To achieve this, the distribution of each topoisomerase was seperated by kmeans clustering analysis into two clusters based on the mean-normalised distribution around the TSSs. In both topoisomerase I and topoisomerase IIβ the cluster seperated the promoters into a 'peak class' and a 'trough class', and in both cases the peak class had significantly higher expression (Figure 3.19a). This is consistent with previous analysis identifying a relationship between expression and topoisomerase enrichment at TSSs (Section 3.2.8.1). The overlap between topoisomerase I and IIβ peaks was established (Figure 3.19b) to identify TSSs that show singular or combined topoisomerase enrichment. The majority of transcription start sites with a peak of topoisomerase have a peak for both topoisomerase I and IIβ (63%), with a notable minority having just topoisomerase I (21%) or topoisomerase IIβ (16%) peaks. The expression level of promoters with a topoisomerase I peak or both a topoisomerase I and IIβ peak are not significantly different, but are significantly enriched compared to TSSs with a topoisomerase IIβ peak only. This suggests that the presence of a peak of topoisomerase IIβ, either together with topoisomerase I or alone, does not correlate with increased expression. To confirm that this is the case, a comparison was made between promoters with a topoisomerase IIβ peak only and the topoisomerase IIβ 'trough class' promoters (Figure 3.19c). Despite a clear difference in the distribution of topoisomerase IIβ around the TSS, there is no significant difference in expression between genes with a topoisomerase IIβ peak only at the TSS and the 'trough class' of promoters. Together this data refutes the model proposed by Kouzine et al. (2013), supporting an alternative model where topoisomerase I is focally enriched in an expression dependent manner, whereas topoisomerase IIβ is enriched at promoters in an expression independent manner.


Topoisomerase I
a)
2.5kb
peak class
trough class
log2 (sample/input)
expression
p < 2.2x10-16
Topoisomerase IIβ
2.5kb
peak class
trough class
log2 (sample/input)
expression
p < 1.2x10-9
b)
Topoisomerase I
Topoisomerase IIβ
231
897
304
expression
ns
p = 1.7x10-9
p = 5.4x10-8

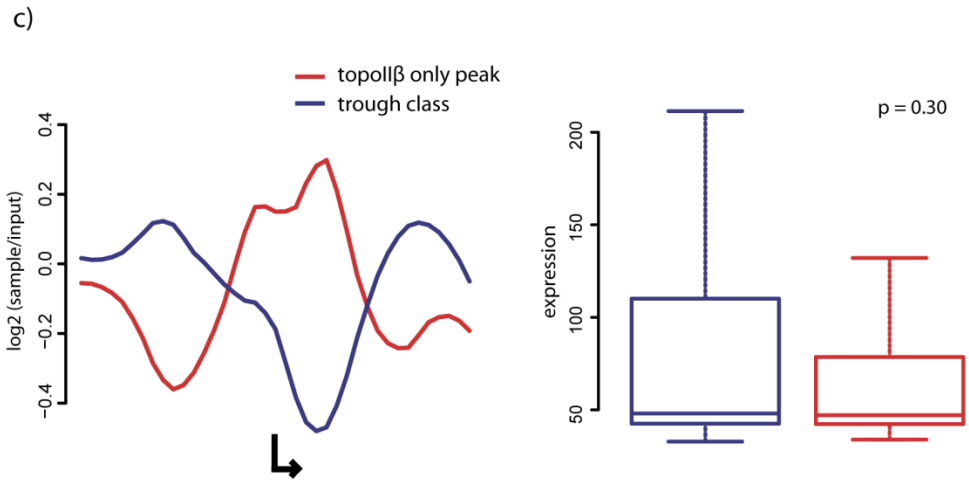


Figure 3.19 Topoisomerase IIβ peaks at TSSs are expression independent. a) Classifying promoters based on topoisomerase distribution. Promoters clustered by k-means analysis with two clusters identify a 'peak class' and 'trough class' for both topoisomerase I and topoisomerase IIβ. A comparison of expression between these clusters supports a relationship between topoisomerase peaks and expression. Significance determined by Student's t-test. b) Classifying promoters based on the presence of topoisomerase I and IIβ peaks. Venn diagram identifies that most promoters have both topoisomerase I and IIβ peaks. Expression analysis of promoters with either topoisomerase I peaks, topoisomerase IIβ peaks or both identifies significantly lower expression in topoisomerase IIβ peak only promoters by Student's t-test. c) Comparing topoisomerase IIβ only promoters with the topoisomerase IIβ trough class identifies no significant difference in expression by Student's t-test.

3.3 Discussion

To provide the first high resolution analysis of topoisomerase I, IIα and IIβ in the human genome, ChIP-chip was performed in a human cell line. During the validation of this ChIP protocol topoisomerases were shown to have a diffuse nuclear distribution in which the majority of topoisomerase is not tightly associated with the chromatin. Despite this, ChIP analysis identified variable enrichment of topoisomerases across the genome.

To identify the importance of topoisomerase enrichments, the distribution of each topoisomerase with respect to the underlying sequence was investigated more thoroughly. An open question regarding topoisomerases was whether they form domain-scale or more focal enrichments in the human genome. An analysis of the distribution of topoisomerases across the loci investigated identified large scale domains of enrichment with a mean span of ~100 kb. These are similar in size to other large-scale regulatory domains in the human genome, including supercoiling domains (Naughton et al., 2013a), topological domains (Dixon et al., 2012) and cytologically determined chromosome loops (Earnshaw and Heck, 1985; Paulson and Laemmli, 1977). Additionally, the co-transcriptional regulation of genes through a 'transcription ripple effect' has been shown to act at scales of 100 kb (Ebisuya et al., 2008), potentially through DNA supercoiling, and could be influenced by topoisomerase distributions at this scale. The domains of topoisomerase I are enriched for GC, TSSs and RNA polymerase II, supporting the observations in *Drosophila* that topoisomerase I is found at transcriptionally active euchromatin (Filion et al., 2010).

The relationship between topoisomerase I and RNA polymerase II is particularaly tight, with a correlation similar to that between different RNA polymerase II antibodies. This supports the direct relationship between topoisomerase I and RNA polymerase II identified in *Drosophila* (Gilmour et al., 1986). Based on RNA polymerase II and topoisomerase I ChIP, the major function of topoisomerase I *in vivo* is likely to be the resolution of transcription induced DNA supercoils. Whether

topoisomerase I performs a similar role in replication cannot be determined from our data and may be difficult to measure by ChIP, due to variability in the position, timing and rate of replication in a population of cells. However, experiments in *Drosophila* indicate that topoisomerase I activity is essential for cell proliferation (Zhang et al., 2000).

The combined enrichment of RNA polymerase II, topoisomerase I and under-wound DNA supercoiling identifies that supercoiled DNA can be maintained in the presence of topoisomerase I *in vivo*. The presence of under-wound DNA has been shown *in vitro* to promote transcription by increasing the rate of transcription intiation at the TSS (Tabuchi and Hirose, 1988). One mechanism that could account for the maintainance of under-wound DNA is the fifty fold higher efficiency of topoisomerase I in the release of over-wound supercoils (Koster et al., 2005). The significant depletion of topoisomerase I in topoisomerase IIβ domains suggests that the presence of over-wound DNA alone is not determining topoisomerase I distribution, as topoisomerase IIβ domains have a more over-wound DNA structure. Therefore, it may be that topoisomerase I is excluded from over-wound or topoisomerase IIβ domains to prevent the relaxation of the repressive DNA structure in these regions.

The distribution of topoisomerase IIα and IIβ domains in the genome are distinct from the topoisomerase I domains. The enrichement of topoisomerase IIα and topoisomerase IIβ is highly similar across domain scale enrichments, supporting at a gross-scale the observations at the MLL gene (Cowell et al., 2012). Topoisomerase II domains are present in AT-rich, gene poor regions of the genome, supporting previous observations in *Drosophila* (Käs and Laemmli, 1992; Miassod et al., 1997) and rat (Sano et al., 2008). This contradicts the study which presents topoisomerase II as the major relaxase in chromatin (Salceda et al., 2006), as the sites with the highest supercoil generation (i.e. most transcriptionally active) are those with the lowest levels of topoisomerase II. Predictions from cytological data suggest that topoisomerase II forms the base of chromatin loops and is enriched in matrix/scaffold attached regions (MARS/SARS) (Earnshaw and Heck, 1985). These MARS/SARS have not been mapped at high resolution, so the length of DNA

associated with them remains unknown. *In vitro* topoisomerase II activity assays in *Drosophila* identify >10 kb enrichments of topoisomerase activity, which they say coincide with SARS. However, the domain-scale enrichments of topoisomerase II observed by ChIP-chip are generally in excess of 100 kb. This indicates that topoisomerase IIα or IIβ do not form points of focal enrichment at defined topological boundaries. Attempts to analyse the distribution of topoisomerases at the boundaries of characterised domains, including LADs (Guelen et al., 2008), supercoiling domains (Naughton et al., 2013a) and topological domains (Dixon et al., 2012), were unsuccessful due to a paucity of boundaries within the regions analysed and the poorly defined nature of these 'boundary' regions. However, it seems likely that the model proposed in cytological studies is incomplete, with some domains being enriched in topoisomerase II over considerable genomic distance. A model whereby over-wound DNA domains are enriched in topoisomerase II is not contradictory to the data presented in these cytological papers, as the proteins associated with the looped DNA regions is removed by high salt to expose the underlying nuclear matrix (Earnshaw and Heck, 1985). Therefore topoisomerase II may be associated with the nuclear matrix and within some loops of the chromatin, perhaps in a continuous domain. To established if this is the case, the precise molecular and genomic characteristics of MARS/SARS must be identified and compared with the distribution of topoisomerase II *in vivo*.

Further characterisation of the topoisomerase II domains identified a relative over-winding of the DNA structure in these regions. Over-wound DNA has been associated with gene repression (Naughton et al., 2013a), supported by the depletion of RNA polymerase II in ChIP experiments. A further potential function of over-wound supercoiling is in the decatenation of chromosomes, a phenomenon that has been observed in prokaryotes (Martínez-Robles et al., 2009) and yeast plasmids (Baxter et al., 2011). The co-localisation of a decatenating DNA structure and the decatenating enzymes topoisomerase IIα and IIβ could indicate that decatenation hotspots occur in gene poor regions of the genome. As decatenation introduces transient double strand breaks into the DNA, which increase the potential for deleterious mutation, restricting this process to gene poor regions could be one mechanism to limit DNA damage in coding regions. The enrichment of

topoisomerase IIα in these regions particularly supports this hypothesis, as it is predominantly expressed in S/G2/M phases of the cell cycle for the decatenation of chromosomes (Woessner et al., 1991). To further establish whether topoisomerase II and over-wound DNA supercoiling work together in the cell-cycle to separate and package chromosomes, topoisomerase ChIP and psoralen-IP should be performed over a time course in synchronised cells.

Mapping topoisomerases by ChIP-chip has identified that topoisomerase primarily form domain scale enrichments that correspond to different regions of the genome, with topoisomerase I enriched in active, under-wound euchromatin and topoisomerase IIα/β enriched in inactive, over-wound repressed chromatin. Another important aim of this chapter was to identify the validity of the model of topoisomerase distribution at promoters proposed by Kouzine et al. (2013). In this model topoisomerase I has a diffuse enrichment upstream of the TSS, in an expression dependent manner, and topoisomerase II is enriched in a focal manner at highly expressed genes. By analysing the distribution of topoisomerase I and IIβ over a 5 kb region centred on the TSS for 2,509 promoters I have established that the proposed model is incorrect. Both topoisomerase I and topoisomerase II form focal distributions at gene promoters for around half of the genes, with no diffuse enrichment observed at the scale of the promoter region. Separating promoters into those with a shared or those with an individual peak of topoisomerase I and II allowed the identification of the relationship between the topoisomerase peak and expression *in vivo*. Genes with a topoisomerase I peaks at the promoter were generally expressed, as suggested by Kouzine et al. (2013). However, genes with a peak of topoisomerase I and II at the promoter did not have a higher expression level than those with only a topoisomerase I peak. Furthermore, those genes with a topoisomerase II peak only had significantly lower expression than either topoisomerase I only or topoisomerase I and II. This contradicts the model by Kouzine et al. (2013) by suggesting that topoisomerase II forms a peak at many gene promoters independent of expression. Other studies have observed topoisomerase II enrichment in non-expressed genes, with a study comparing gene expression in WT and topoisomerase IIβ KO rat brains suggesting only ~2% of genes were dependent on topoisomerase IIβ for their regulation (Lyu et al., 2006). Therefore, the model

proposed by Kouzine et al. (2013) is inappropriate as a general model for topoisomerase function, but it is possible that it applies to a small subset of genes. Based on the extensive topoisomerase ChIP-chip data generated, a model in which topoisomerase I forms an expression depdendent peak of enrichment at the TSS whereas topoisomerase II is present at many promoters in an expression independent capacity is more appropriate. This model is further supported by the ratchet like mechanism of topoisomerase I, which releases more DNA supercoils per reaction when the tension in the DNA is higher (Koster et al., 2005). Topoisomerase II, on the other hand, is limited to the release of two supercoils per reaction independent of DNA supercoil intensity. Therefore, the role of topoisomerase I activity in the release of DNA supercoils at expressed gene promoters is clear, whereas the mechanism of topoisomerase II remains unclear.

The aim of this chapter was to provide a comprehensive understanding of topoisomerases in the human genome. Through a ChIP approach topoisomerases have been shown to form diffuse distributions over large domains and more focal distributions at transcription start sites. This suggests that both topoisomerase I and II proteins function at different scales in the human genome. Topoisomerase I is associated with active regions and at the promoters of active genes, maintaining an under-wound structure through a diffuse distribution at a ~100 kb scale and a focal distribution at transcription start sites. Topoisomerase II proteins are found in gene poor regions and at the transcription start sites of around half of the genes analysed in an expression independant manner. The function of topoisomerase II proteins in these locations is uncharacterised *in vivo*. One interesting possibility is that topoisomerase II enriched regions are decatenation hotspots, with further work necessary to test this hypothesis. The distribution of topoisomerases at transcription start sites does not follow the model proposed by Kouzine et al. (2013) and a new model is proposed in which topoisomerase I is the major factor in the relief of transcription dependent supercoils *in vivo*.

4. DNA supercoiling at gene promoters

4.1 Introduction

The distribution of topoisomerases (Chapter 3) and the identification of DNA supercoil domains (Naughton et al., 2013a) in the human genome raises questions about the role of DNA supercoiling in regulating gene expression *in vivo*. A number of studies have indicated that the DNA supercoiling primarily influences expression through the modification of gene promoter structure (Mizutani et al., 1991a, 1991b; Tabuchi and Hirose, 1988). For example, an *in vitro* experiment that separates the transcription of the *Bombyx mori* fibroin gene into a) initiation complex formation, b) conversion to the elongation complex and c) subsequent elongation, identifies that only the initiation complex formation is positively influenced by under-wound DNA supercoiling (Tabuchi and Hirose, 1988). The identification of a focused enrichment of under-wound DNA supercoils at gene promoters in human, fly and hamster supports this mechanism *in vivo* (Jupe et al., 1993; Ljungman and Hanawalt, 1992, 1995). Furthermore, two recent analyses of several hundred human promoters have identified that an 'average' gene promoter has a transcription dependent under-wound structure (Kouzine et al., 2013; Naughton et al., 2013a). To better understand the role of DNA supercoiling at human gene promoters a much larger-scale analysis is required, in which promoters can be separated by known and novel structural properties.

Two general approaches have been developed to determine the distribution of DNA supercoiling, taking advantage of the characteristic transitions in DNA twist and writhe. One of these methods utilises the differential migration of relaxed and supercoiled plasmid DNA through gradients/gels and separates them based on plasmid structure, whereas the second method utilises a psoralen based probe that preferentially intercalates into under-wound DNA. Using variations on these two techniques, many *in vitro* and *in vivo* characteristics of supercoiled DNA have been identified.

Early studies of DNA plasmids by sucrose gradient sedimentation identified that different structures exist in a sample containing only plasmids of equal molecular weight. The nature of this structural difference was identified by Vinograd et al. (1965), who showed that a single-strand nick caused the sedimentation of a single structure and that plasmid DNA usually had a 'twisted circular structure', subsequently named supercoiled DNA. The gross structural difference between relaxed and supercoiled DNA plasmids measured in these sucrose gradient experiments are now utilised in several assays for DNA supercoiling. In each case supercoiled plasmids move through the gradient/gel more rapidly than relaxed plasmids, due to their more compact structure. In agarose gel electrophoresis the difference between relaxed and supercoiled plasmids is particularly clear, with each topoisomer forming a distinct enrichment to give a 'ladder' of DNA molecules with different levels of supercoiling (Figure 4.1a). One dimensional (1D) agarose gel electrophoresis clearly displays the difference between supercoiled and relaxed plasmids, but a modification of this protocol in a second dimension allows the resolution of a wider range of topoisomers including the identification of positively and negatively supercoiled plasmids. In 2D agarose gel electrophoresis plasmids are initially separated through a gel as in 1D agarose gel electrophoresis. The gel is then soaked in a buffer containing the intercalating agent ethidium bromide, which introduces over-wound supercoils into the plasmids, and run at 90° to the original electrophoresis separating the plasmids in a second dimension (Figure 4.1b). These techniques have been used extensively to model the properties of supercoiled DNA *in vitro*, but are limited to plasmid systems. However, an elegant adaptation of the cre-lox system has allowed the characterisation of DNA supercoiling at a limited number of sites *in vivo* (Kouzine et al., 2008). This study aimed to identify the structural properties of the FUSE sequence under *in vivo* DNA supercoiling conditions. To do this a construct containing loxP sites flanking the region of interest was stably transfected into human cells and the cre-lox system used to excise DNA minicircles in the *in vivo* supercoiled state for analysis by 2D gel electrophoresis. Using this technique it was established that under-wound DNA upstream of promoters increases with transcription and that supercoiling can cause the FUSE element to adopt a non-B DNA conformation that recruits specific

transcription factors. This clearly demonstrates the importance of DNA supercoiling *in vivo*, but is not easily scaled to study large portions of the genome.

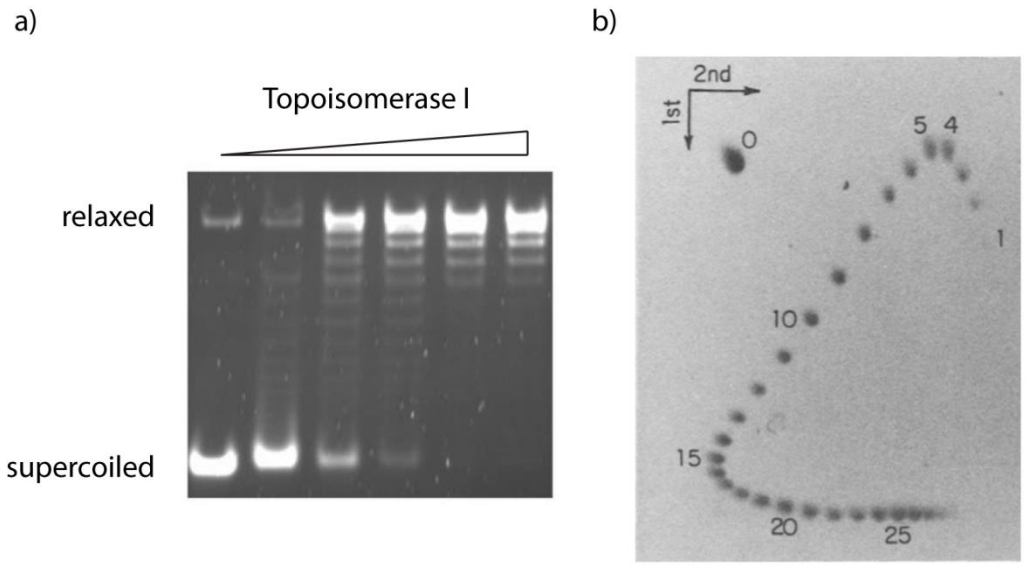


Figure 4.1 Differential migration of relaxed and supercoiled DNA plasmids. a) Agarose gel electrophoresis of DNA in a supercoiled state and progressively relaxed with topoisomerase I clearly identifies topoisomers. Image from Li et al. (2005). b) Two dimensional (2D) agarose gel electrophoresis. Topoisomer 15 is relaxed DNA, topoisomers 14 to 1 are more over-wound and topoisomers 16 to 28 are more under-wound. Image from Wang et al. (1983).

To characterise the genome-wide distribution of DNA supercoiling *in vivo*, techniques have been developed based on the relative intercalation of psoralen. There is a linear relationship between DNA supercoiling and psoralen intercalation, with an increase from over-wound, through relaxed to under-wound DNA (Bermúdez et al., 2010). This preference occurs because the intercalation of psoralen into DNA induces a slight over-wound twist to the double helix, which is more energetically favourable on an under-wound template. Once intercalated a psoralen molecule can be covalently cross-linked to the DNA by UV irradiation and the relative enrichment determined on purified DNA. Initial analysis focussed on the difference between prokaryotic and eukaryotic genomes, concluding that the genome of *E. coli* is maintained in an unrestrained under-wound state whereas eukaryotic genomes have no net unrestrained DNA supercoiling (Sinden et al., 1980). Subsequent analyses at specific promoters, enhancers and genomic loci have identified that there is in fact variability in DNA supercoiling across eukaryotic genomes, including yeast, fly and human (Bermúdez et al., 2010; Jupe et al., 1993; Kouzine et al., 2013; Ljungman and Hanawalt, 1992, 1995; Matsumoto and Hirose, 2004; Naughton et al., 2013a). Two recent papers have adapted the use of psoralen for the identification of DNA supercoiling over large regions of the genome by microarray analysis. To enrich for psoralen bound DNA, Kouzine et al. (2013) UV cross-linked psoralen to DNA stands and, after denaturation, digested the unbound single stranded DNA. Our lab has developed a second approach to enrich for psoralen, using a biotinylated trimethylpsoralen (bTMP) which can be selectively immunoprecipitated using streptavidin coated beads (Naughton et al., 2013a). I have used this approach to investigate the distribution of DNA supercoiling at human gene promoters to understand the role of DNA supercoiling in regulating promoter structure and function

Psoralen is a tricyclic compound, composed of a furan ring and a coumarin, that forms covalent bonds with pyrimidine bases upon photo-crosslinking (Cimino et al., 1985). A number of derivatives of psoralen have been isolated either as a natural plant product or through chemical synthesis, including isopsoralen, 8-methoxypsoralen, 4,5’ ,8-trimethylpsoralen, 4’ –hydroxymethyl-4,5’ ,8-trimethylpsoralen and 4’aminomethyl-4,5’ ,8-trimrthylpsoralen. The most

commonly used psoralen molecule for the investigation of DNA structure is 4,5',8-trimethylpsoralen (Figure 4.2a) (TMP). TMP is a cell permeable planar molecule that intercalates between base pairs in the DNA double helix and forms stable photo-crosslinks with pyrimidine nucleotides upon exposure to near UV light (Cech and Pardue, 1977). TMP can form mono-adducts or inter-strand cross-links in the DNA double helix between the 4',5' furan double bond and/or the 3,4 pyrone double bond of the TMP molecule (Figure 4.2a) and the 5,6 double bond of cytosine or thymine (Figure 4.2b, available bond in red, unavailable bind in blue) (Kanne et al., 1982). Despite available interaction sites for both thymine and cytosine nucleotides, a number of studies have indicated that TMP has a preference for thymine cross-links (Esposito et al., 1988; Kanne et al., 1982; Song and Ou, 1980). Whether this preference results from differences in nucleotide structure or the structure of helices containing cytosines is currently unknown. Further analysis of the sequence preference of TMP identified a complicated and unpredictable relationship, with a preference for 5'TA over 5'AT, a strong influence of flanking bases up to 3 bp either side of the interaction site and potential long range effects over tens of base pairs (Esposito et al., 1988). The clear influence of the local sequence context on TMP-DNA interactions suggests that the DNA helical structure is important for TMP binding independent of nucleotide structure.

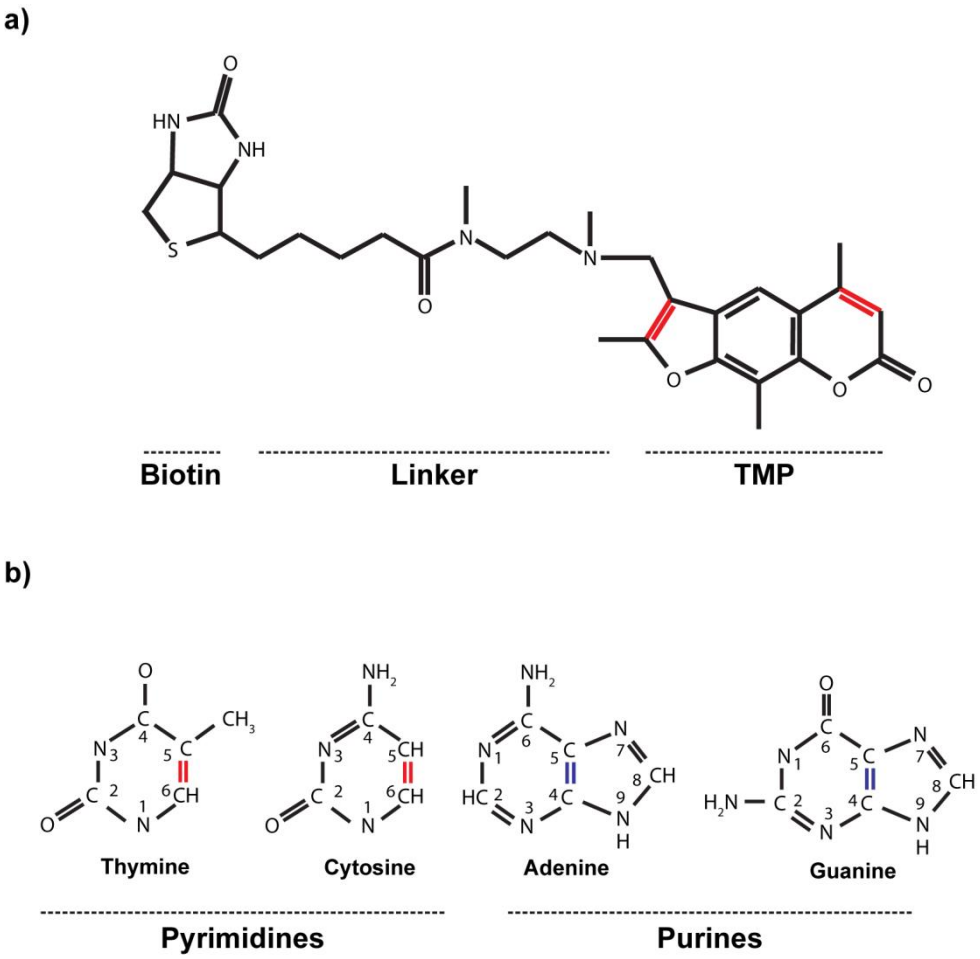


Figure 4.2 bTMP and nucleotide structures indicate potential cross-linking sites. a) The structure of bTMP. The molecule is made up of a biotin, a linker and a TMP. BTMP forms adducts with nucleotides at the 4',5' furan double bond and/or the 3,4 pyrone bond, both marked in red. b) The structure of the nucleotides in DNA. BTMP forms adducts with the 5,6 double bond, which is available in pyrimidines (marked red) and unavailable in purines (marked blue).

The intercalation of TMP into the DNA introduces over-wound DNA supercoils into the DNA, in a mechanism similar to other intercalating agents including ethidium bromide and chloroquine (Bates and Maxwell, 2005). As the over-winding of an under-wound DNA helix towards a more relaxed state is more energetically favourable than further over-winding an already over-wound helix, the intercalation of TMP favours an under-wound DNA helix. This relationship between DNA supercoiling and TMP intercalation has been shown to be linear in a plasmid system (Bermúdez et al., 2010), indicating that TMP binding gives a good representation of the underlying DNA supercoiling. Using TMP derivatives a number of groups have identified regions of relative under-/over- winding in the genomes of model organisms (Bermúdez et al., 2010; Jupe et al., 1993; Ljungman and Hanawalt, 1995; Matsumoto and Hirose, 2004) and human cell lines (Kouzine et al., 2013; Ljungman and Hanawalt, 1992; Naughton et al., 2013a). Together with *in vitro* studies which suggested that under-wound DNA increases transcription efficiency and promotes the formation of a pre-initiation complex (Mizutani et al., 1991a, 1991b; Tabuchi and Hirose, 1988), it became apparent that DNA supercoiling at gene promoters had the potential to regulate gene expression in eukaryotes. However, a genome-wide analysis of DNA supercoiling at human gene promoters has not been examined.

Two recent publications have taken the first steps in understanding the general principles of DNA supercoiling at transcription start sites in human cells. Using a novel bTMP pull-down approach, our lab performed a meta-analysis on the promoters of 584 genes and identified a topoisomerase and transcription dependent peak of under-wound DNA at gene promoters, which was particularly enriched at expressed genes (Naughton et al., 2013a). A second study identified a similar under-wound structure in the absence of transcription inhibitors through a meta-analysis of 445 gene promoters (Kouzine et al., 2013). These studies identify transcription dependent differences in DNA supercoiling at gene promoters but, due to their small-scale, are unable to identify DNA supercoiling promoter sub-types.

To establish structural and functional classes of gene promoter based on DNA supercoiling, I have performed a more extensive analysis of gene promoters using the bTMP pull-down approach. In this technique a cell-permeable bTMP molecule

(Figure 4.3) (Saffran et al., 1988) is added to the culture media and after a short incubation period the cells are cross-linked by irradiation at 365 nm. Cells are then lysed, the DNA is fragmented by sonication, purified and immunoprecipitated using streptavidin coated magnetic beads. The enriched psoralen-bound DNA and corresponding inputs are amplified, labelled and hybridised to 2 colour microarrays covering regions of interest. Using this technique, my study will characterise the supercoiling properties of gene promoter using genome wide promoter microarrays and to relate this to topoisomerase binding (Chapter 4) and other genomic features including transcription and base composition. This analysis aims to identify novel properties of gene promoters and provide a comprehensive understanding of promoter DNA supercoiling in the human genome.

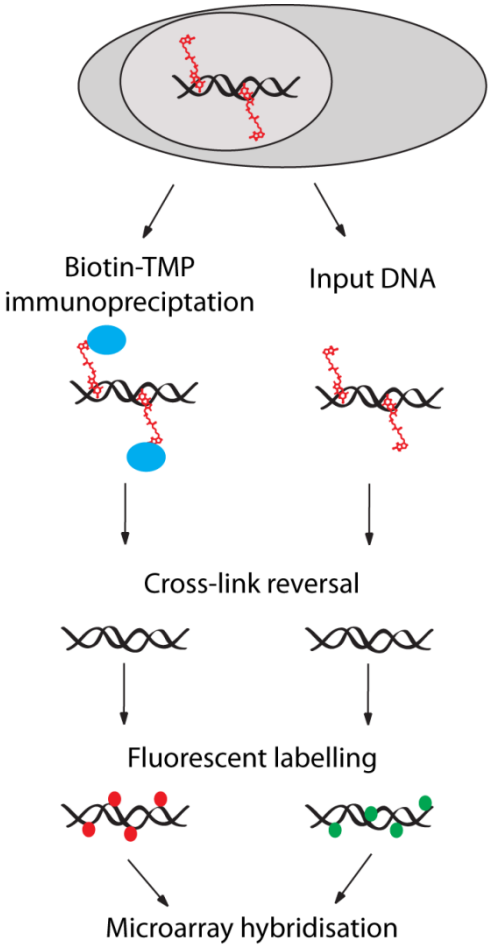


Figure 4.3 bTMP immunoprecipitation. Cells were incubated with bTMP (red structure) and cross-linked by irradiation at 365 nm. Cells were lysed and the DNA fragmented by sonication and purified. To enrich for under-wound DNA samples were immunoprecipitated using streptavidin magnetic beads (blue circles) and as a control non-immunoprecipitated input samples were retained. The immunoprecipitated and input samples were heat treated in formamide to reverse the bTMP cross-link. These DNA samples were amplified and labelled (red and green circles) before hybridisation to Nimblegen 2.1M promoter arrays (see materials and methods).

4.2 Results

4.2.1 bTMP pull-down validation

4.2.1.1 Synthesis and characterisation of the psoralen molecule

In order to characterise DNA supercoiling at gene promoters bTMP was synthesised in conjunction with the group of Mark Bradley at the University of Edinburgh Chemistry Department. The bTMP molecule was based on the molecule described in Saffran et al. (1988) and synthesised from trioxsalen as previously described (Naughton et al., 2013a; Saffran et al., 1988). To confirm the purity of the bTMP an aliquot was analysed by high performance liquid chromatography mass spectrometry (HPLC-MS), identifying a single prominent peak by HPLC (Figure 4.4a), which together with NMR and ELSD data (data not shown) indicates a high level of sample purity. HPLC-MS on this peak identifies two clear peaks at 555.1/556.3/557.2 and 577.3/578.3/579.1 (Figure 4.4b). The first peak corresponds to bTMP (molecular weight 554 Da) plus one/two/three protons whilst the second peak corresponds to bTMP plus two sodium and one/two/three p. There are a number of peaks which do not correspond to the major constituent chemicals of the synthesis reaction, including trioxsalen (MW 228.24), 4'-(chloromethyl)trioxsalen (MW 276.7), 1,2-dimethyldiaminoethane (MW 88.15) or NHS biotin (MW 341.38). It is difficult to draw conclusions about these other masses, but it is clear that the synthesis reaction has produced bTMP for subsequent use in the identification of promoter DNA supercoiling.

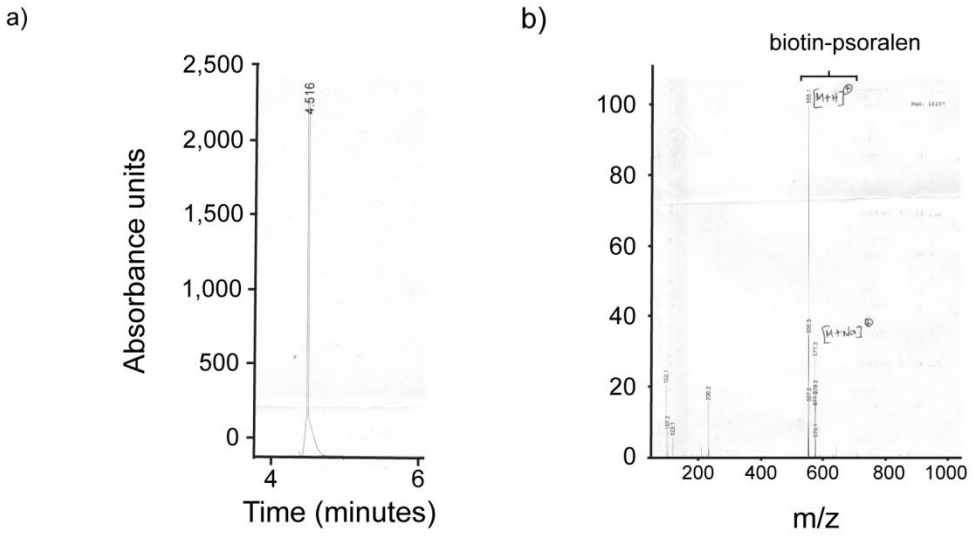


Figure 4.4 HPLC-MS confirms bTMP synthesis. a) HPLC identifies a single strong peak in the bTMP synthesis product. b) MS identifies bTMP within the HPLC peak.

4.2.1.2 bTMP preferentially binds thymine *in vitro*

Psoralen molecules have a reported binding preference for the nucleotide thymine, despite an available binding site in both cytosine and thymine (Figure 4.2). In addition, several structural features of our bTMP molecule (Naughton et al., 2013a) could influence the reaction with DNA compared to other psoralens, including the charged linker region and increased molecular weight. To establish the sequence preference of our bTMP a photo-crosslinking experiment was performed on poly-d(AT) and poly-d(GC) oligonucleotides. To establish the size of the commercial poly-purine:pyrimidine oligonucleotides DNA samples were run on a TBE agarose gel prior to photo-crosslinking (Figure 4.5a). The poly-d(A)T oligonucleotides had a length of 500-1200 bp whereas the poly-d(GC) had a length ~400 bp, therefore each oligonucleotide has the potential to bind many bTMP molecules. Photo-crosslinking experiments between these DNA fragments and bTMP identifies a reaction with poly-d(AT), but not with poly-d(GC) (Figure 4.5a). Therefore, under these experimental conditions bTMP binds only thymine at detectable levels within this sequence context.

The identification that DNA sequence can influence psoralen intercalation tens of base pairs from the site of interaction suggests that the helical structure of DNA is important factor in intercalation frequency (Esposito et al., 1988). It is well established that poly-d(GC) oligonucleotides have a propensity to form non-B form DNA structures, including Z form structures (Thamann et al., 1981), which may influence bTMP intercalation. To establish under a broad range of sequence contexts the binding preference of bTMP a photo-crosslinking reaction was performed on fragmented human genomic DNA (Figure 4.5b), followed by digestion to mononucleotides and analysis by HPLC-MS. This experiment was performed once and identified a peak corresponding to bTMP bound thymine (MW 796 Da) in the photo-crosslinked sample (Figure 4.5b), but not in the non-crosslinked control. No peak corresponding to bTMP bound cytosine (MW 781 Da) was identified in either the cross-linked or non-crosslinked sample. Therefore under our experimental conditions *in vitro* bTMP binds detectably to thymine, but not to cytosine. This supports the preferential binding of psoralen to thymine observed in previous studies,

but cannot exclude the binding of bTMP to cytosine at levels below the threshold of detection.

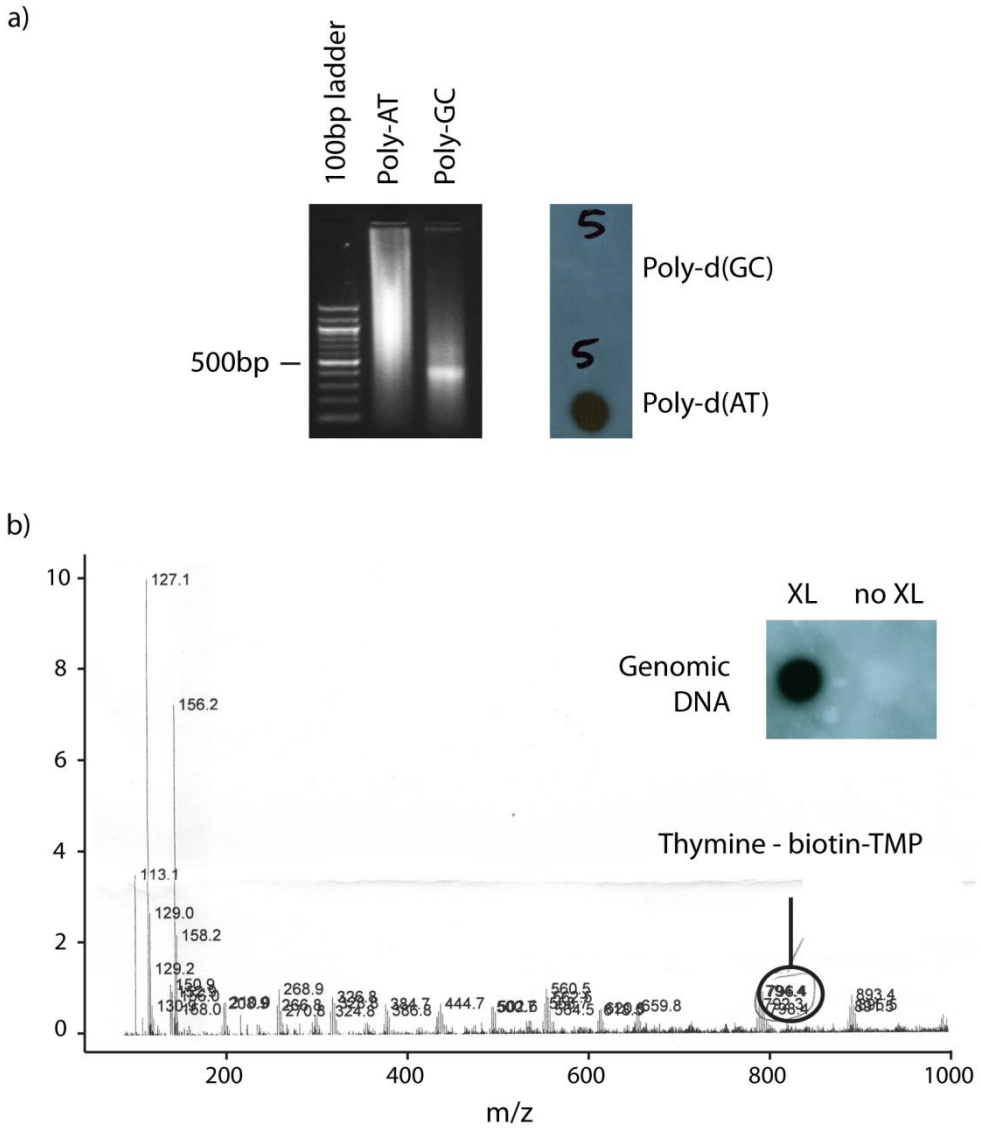


Figure 4.5 bTMP binds to thymine. a) bTMP preferentially binds to poly-AT. Left panel shows the size distribution of poly-AT and poly-GC oligonucleotides by agarose gel electrophoresis. Right panel shows a dotblot probed for biotin following a bTMP photo-crosslinking reaction for each oligonucleotide. b) bTMP binds thymine in genomic DNA. HPLC-MS for digested genomic DNA cross-linked to bTMP. Inset shows a dotblot of the bTMP bound DNA sample, with and without cross-linking (XL) by irradiation at 365 nm, identified with avidin-HRP.

4.2.1.3 Binding of bTMP to A-form and B-form DNA helices

In an attempt to characterise the DNA structure preference of bTMP further, photo-crosslinking experiments were performed on DNA oligonucleotides with different helical structures. Previous studies have analysed the structure of specific oligonucleotide sequences by X-ray crystallography and established that different sequences adopt distinct helices (Hays et al., 2005). These helices can have markedly different structures, which may expose or conceal psoralen binding sites in the DNA helix (Section 1.1.1.2). To establish if TMP based molecules can differentiate between stable helices, in addition to their preferential intercalation into under-wound DNA, a photo-crosslinking experiment was conducted on oligonucleotides with an A, B or AB intermediate structure (for sequence see Materials and Methods Section 2.2.6.1). Photo-crosslinking of bTMP to these DNA sequences identified preferential binding to the AB intermediate, but no detectable binding to the specific A or B forms of DNA (Figure 4.6). The AB intermediate oligonucleotide sequence has four thymines compared with two for the A and B form molecules, which may account for this difference in bTMP binding independent of helical structure. Crystal structure analysis has not identified a suitable oligonucleotide sequence with a defined A or B form helix that contains more than two thymines (Hays et al., 2005). Therefore, under these experimental conditions it was not possible to determine whether bTMP preferentially binds A, B or A/B form DNA helices independent of thymine levels.

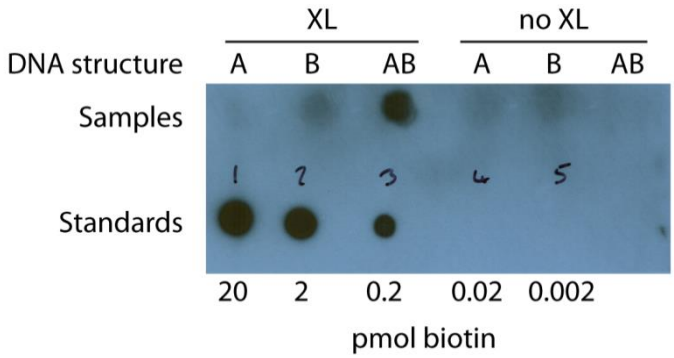


Figure 4.6 bTMP photo-crosslinking to A, B and AB oligonucleotides. Dotblot experiment in which 'samples' are DNA purified from A, B and AB oligonucleotides incubated with bTMP, with and without photo-crosslinking. 'Controls' are DNA oligonucleotides with a single modified base containing a biotin. The concentration therefore represents the concentration of DNA oligonucleotides and of biotin in each dot.

4.2.1.4 bTMP binding *in vivo*

To confirm that bTMP is cell permeable and to establish the crosslinking frequency in the genome, a photo-crosslinking experiment was performed in Retinal Pigmented Epithelial (RPE1) cells. Cells were incubated with bTMP and UV irradiated, followed by DNA isolation, purification. Dot-blot analysis with a streptavidin conjugated horseradish peroxidase showed that bTMP bound to DNA following UV irradiation *in vivo* (Figure 4.7), in agreement with previous studies (Naughton et al., 2013a; Saffran et al., 1988). The cross-linking frequency, calculated with respect to the biotin-oligonucleotide standards, is around one bTMP per 900 bp of double stranded DNA. At this level of cross-linking our lab had previously identified peaks of bTMP at transcription start sites (Naughton et al., 2013a), indicating that this cross-linking frequency is sufficient to analyse different promoter structures.

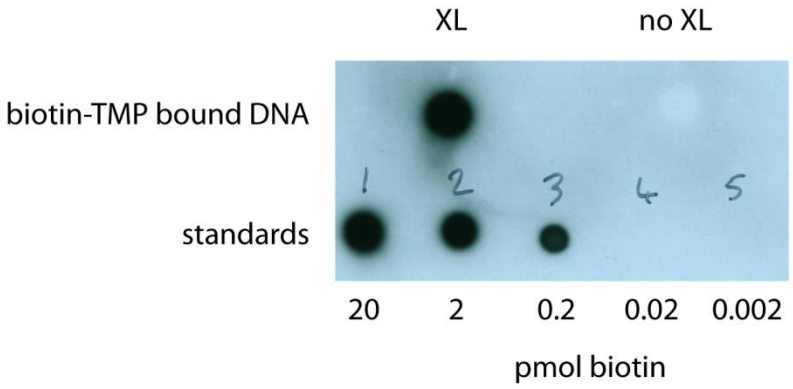


Figure 4.7 bTMP binding *in vivo*. Dotblot for DNA isolated from a bTMP photo-crosslinking experiment in RPE1 cells. Standards are DNA oligonucleotides with a single modified base containing a biotin. The standards show the concentration of DNA oligonucleotides and therefore biotin in each spot.

4.2.2 bTMP pull-down

4.2.2.1 Hybridisation of bTMP enriched DNA to genome wide promoter microarrays

To identify the structural properties of promoters genome wide, the bTMP immunoprecipitation samples from Naughton et al. (2013) were re-labelled and hybridised to Nimblegen 2.1M promoter arrays at the VUMC microarray facility (for bTMP-IP experimental procedure see Materials and Methods Section 2.6.5). Samples included non-treated cells ('control') and those treated with the transcription inhibitor α-amanitin for 5 hours ('α-amanitin') followed by a 3 hour wash-out ('wash-out'). As a control for DNA sequence bias bTMP immunoprecipitation on naked genomic DNA ('genomic') was also hybridised to microarrays. Together this data enables the investigation of promoter DNA structure under 'steady state' conditions and following transcription perturbation and recovery, which disrupts promoter DNA structure *in vivo* (Kouzine et al., 2013; Naughton et al., 2013a). Following sample labelling and array hybridisation, the VUMC facility provided the raw data, which was processed as outlined in Section 3.2.3.1 for quality control and signal normalisation. Subsequent analysis was performed on this normalised dataset consisting of control, α-amanitin, α-amanitin wash-out and genomic bTMP immunoprecipitation samples.

4.2.2.2 bTMP immunoprecipitation of genomic DNA identifies a subtle thymine preference

To identify bioinformatically whether the bTMP thymine preference observed at the scale of single nucleotides influences the pull-down efficiency of genomic DNA fragments (~ 300 bp) of different nucleotide composition, the relative enrichment of bTMP was compared to GC percentage for all DNA probes on the microarray. A subtle negative relationship was observed between GC% and bTMP enrichment (Figure 4.8). However, there is not a strong depletion for bTMP, even at very high GC% (i.e. low thymine %). This result is surprising, given the *in vitro* experiments discussed previously, but has been observed previously in our lab. We interpret this

result as a reflection of a poor understanding of the mechanism of bTMP binding, beyond an association with under-wound DNA. For example, the presence of a small number of thymine residues on a 300bp fragment of DNA may be sufficient for the binding of bTMP. This could occur if the binding of one bTMP molecule precluded the binding of more molecules to a fragment. In which case, the proportion of thymine nucleotides on a fragment would have a lower than expected influence on the binding of bTMP. Figure 4.8 shows that under the experimental conditions of the bTMP pull-down there is not a general GC sequence bias, therefore subsequent analysis is based on this assumption. A subsequent aim, beyond the scope of this thesis, is to test this assumption more thoroughly through additional experiments and analysis of recently published datasets (see Discussion).

An additional potential bias identified by previous studies is the complex relationship between local nucleotide sequence and psoralen binding, where nucleotides several base-pairs away can influence binding dynamics (Esposito et al., 1988). To account for this complex bias and to characterise the combined influence of chromatin and *in vivo* DNA supercoiling on DNA structure, subsequent analysis of *in vivo* bTMP distributions is corrected for the bTMP enrichment on genomic DNA.

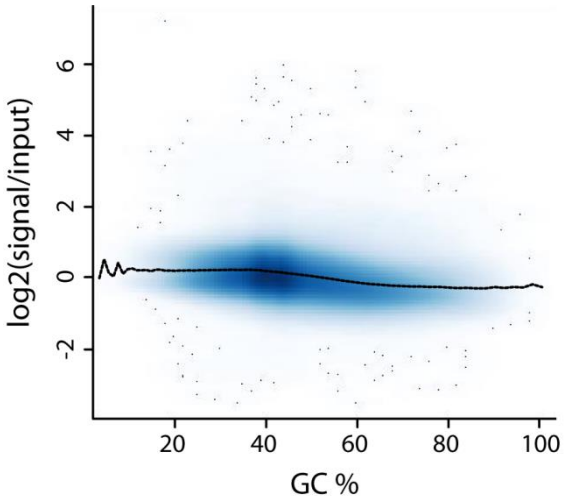


Figure 4.8 bTMP has limited thymine sequence preference in sonicated genomic DNA. Scatterplot showing the relationship between GC% and bTMP enrichment, data taken from a bTMP immunoprecipitation experiment on genomic DNA. The black line represents the median signal value binned for each GC% integer.

4.2.3 bTMP binding identifies distinct DNA supercoiling structures at human gene promoters

bTMP binding is influenced by DNA supercoiling, DNA sequence and base composition *in vivo*. To identify the structural properties of human gene promoters in the context of chromatin, independent of base composition and DNA sequence, that may have unknown structural parameters, the *in vivo* bTMP pull-down probes were each corrected for sequence by subtracting the genomic DNA bTMP pull-down values. This correction permits the direct interrogation of DNA supercoiling at steady-state ('control') and following transcription inhibition ('α-amanitin') and recovery ('wash-out').

4.2.3.1 Transcription dependent peak of DNA supercoiling at TSS

To identify if the transcription dependent peaks of under wound DNA, previously observed at a sub-set of gene promoters (Kouzine et al., 2013; Naughton et al., 2013a), is a general property of promoters genome-wide the median distribution of bTMP was identified across 20,631 promoters. The 'Control' dataset has an enrichment for under-wound DNA around the promoter regions (Figure 4.9), consistent with previous studies. Surprisingly, at the transcription start site (TSS) itself there is a dip of bTMP binding, identifying an over-wound DNA region. This over-wound structure at the TSS may have been missed in previous analyses, due to the application of smoothing algorithms, the investigation of a larger domain around the TSS or the relatively small number of TSSs assayed. This distribution is, however, consistent with an analysis performed on the promoters of expressed genes across the chromosome 11 array (Figure 3.18). Therefore, the DNA structure at an 'average' human promoter is under-wound 1 kb up-/down- stream of the TSS and over-wound DNA at the TSS.

To investigate the role of transcription in the organisation of DNA supercoiling at gene promoters, the median distribution of bTMP binding was identified for promoters genome wide following transcription inhibition with α-amanitin.

Transcription inhibition remodels promoter DNA supercoiling, which adopts a similar pattern to that of genomic DNA (dotted line through zero) (Figure 4.9). This re-organisation appears less substantial than that observed by Naughton et al. (2013), although this is likely to be a technical rather than biological observation. In the present analysis the investigation of bTMP distribution is limited to promoter regions only, due to the design of the microarray, therefore the within array normalisations only covers sequences around promoter regions. In the analysis by (Naughton et al., 2013a) the bTMP binding was assayed across whole megabase scale loci and a sub-set of the normalised data taken for promoter analysis. This means that the generally under-wound structure observed across the promoters in Naughton et al. (2013a) cannot be identified in a specific promoter array analysis, as the normalisation corrects half of the data to be relatively over-wound. Therefore, in the promoter array analysis the shift of the α-amanitin sample to be more under-wound upstream of the TSS and more over-wound around the TSS signifies a strong remodelling of promoter DNA structure toward a more genomic like structure (Figure 4.9), similar to previous observations (Naughton et al., 2013a). This data supports at a genome-wide scale the transcription dependent under-wound structure of gene promoter regions and identifies a novel over-wound DNA structure at the TSS.

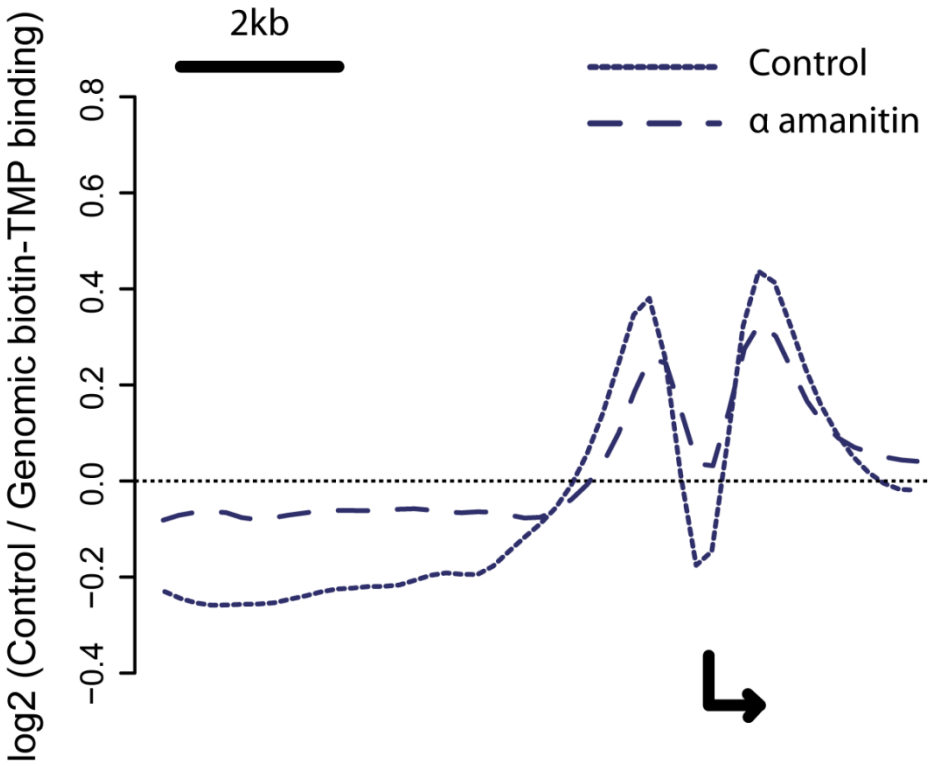


Figure 4.9 Promoters genome-wide have a transcription dependent DNA structure. The median distribution of bTMP in the region 7 kb upstream to 3 kb downstream of TSSs genome-wide for control and α-amanitin samples. Corrected for genomic DNA bTMP enrichment, which is represented by the dotted line through zero.

4.2.4 DNA supercoil distribution is influenced by the presence of a CpG island

The stable structure at the TSS identified in the distribution of bTMP binding across gene promoters may result from a DNA sequence with unusual structural properties. The most common sequence feature at TSSs is the CpG island, a region of elevated CpG density extending for ~1 kb surrounding the TSS of most human genes (Bird, 1986; Bird et al., 1985; Jones, 2012). Furthermore, CpG islands are known to form unusual DNA structures (Lipps and Rhodes, 2009; Rich and Zhang, 2003) and have a less flexible DNA conformation than AT rich regions (Bates and Maxwell, 2005). To identify how CpG islands influence DNA supercoil distribution, promoters were ranked based on their probability of having a CpG island and the relative distribution of bTMP binding identified. A probabilistic approach to CpG island classification has been developed (Irizarry et al., 2009; Wu et al., 2010) to replace the arbitrary definition proposed by Gardiner-Garden and Frommer (1987), in which a CpG island is defined as having an observed-to-expected CpG ratio greater than 0.6 and a GC content greater than 0.5. Using the hidden Markov model (HMM) based approach, promoters were ranked based on the probability of enriched CpG ratio and GC content when compared to the surrounding sequence context. Plotting a heatmap of ranked promoter supercoil distribution identifies a striking pattern in the data (Figure 4.10), in which promoters with the highest CpG island probability have strong peaks of enrichment up-/down- stream of the TSS and a strong over-wound structure at the TSS whilst promoters with the lowest CpG island probability have no clear DNA supercoil distribution. The dotted line through the heat plot represents the cut-off used for a canonical list of CpG islands, similar to those used in the UCSC genome browser which are based on Gardiner-Garden and Frommer (1987) algorithm (Irizarry et al., 2009). This line clearly differentiates the DNA supercoil data into regions with a CpG island-like structure and those with a non-CpG island structure. To establish how the distribution of DNA supercoils across promoters relates to the peak-trough-peak distribution seen in the median distribution, inflections were established for the median distribution of 100 promoter bins in the ranked dataset (Figure 4.10). Inflections are positions in the distribution that form a peak (yellow circles) or trough (blue circles). The inflections plot corroborates the heatmap by

showing that CpG island promoters have a peak-trough-peak distribution, similar to that of the genome-wide median distribution, whereas non-CpG island promoters have a wide peak at the promoter and in most cases show no trough at the TSS. A second striking observation from the turnpoints plot is the regularity of DNA structure at gene promoters, compared to the high variability in structure up- and down- stream. Together, this data supports a model whereby gene promoters have a distinct DNA structure compared to surrounding DNA and that this structure is different for CpG and non-CpG island promoters.

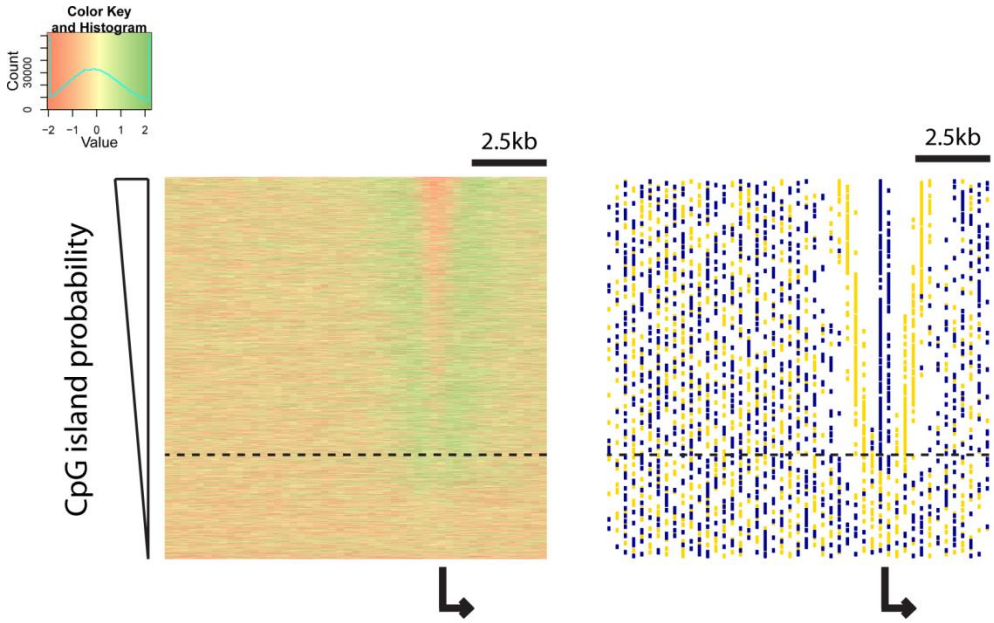


Figure 4.10 DNA supercoil distribution differs for CpG and non-CpG island promoters. The left panel shows a heatmap for bTMP distribution at promoters genome wide, ranked on the probability of having CpG island properties. The colours represent log2(control-genomic) bTMP enrichment from depleted (red) to enriched (green) following the 'colour key and histogram'. The right panel shows the inflections for the same promoter distribution established from the median distribution of 100 promoter bins. Yellow dots represent upward inflections (peaks) and blue dots represent downward inflections (troughs). The dotted line through both panels represents the probability cut-off for canonical CpG island promoters (Irizarry et al., 2009), with all promoters below the line being non-CpG island promoters.

The classification of gene promoters by DNA supercoiling identified a clear separation into CpG island and non-CpG island promoters. To further investigate DNA supercoiling properties the median distribution of bTMP binding was analysed at promoters with CpG islands (12,786 TSSs) compared to those without (7,843 TSSs) (Figure 4.11). Separating the promoters based on the presence of a CpG island shows that the over-wound TSS is associated with CpG island promoters but is absent from non-CpG island promoters. The striking distribution of DNA supercoiling at CpG island promoters identifies a substantial difference in structure between the CpG island itself and the DNA 1 kb up-/down- stream. The distribution of DNA supercoils at non-CpG island promoters is more gradual, with an over-wound upstream region becoming progressively more under-wound towards the TSS. This difference in promoter structure between CpG island and non-CpG island promoters could be an important signal for genetic regulation. To address this subsequent analysis will focus on how the DNA structure of CpG and non-CpG island promoters relates to gene expression, sequence, function and regulation.

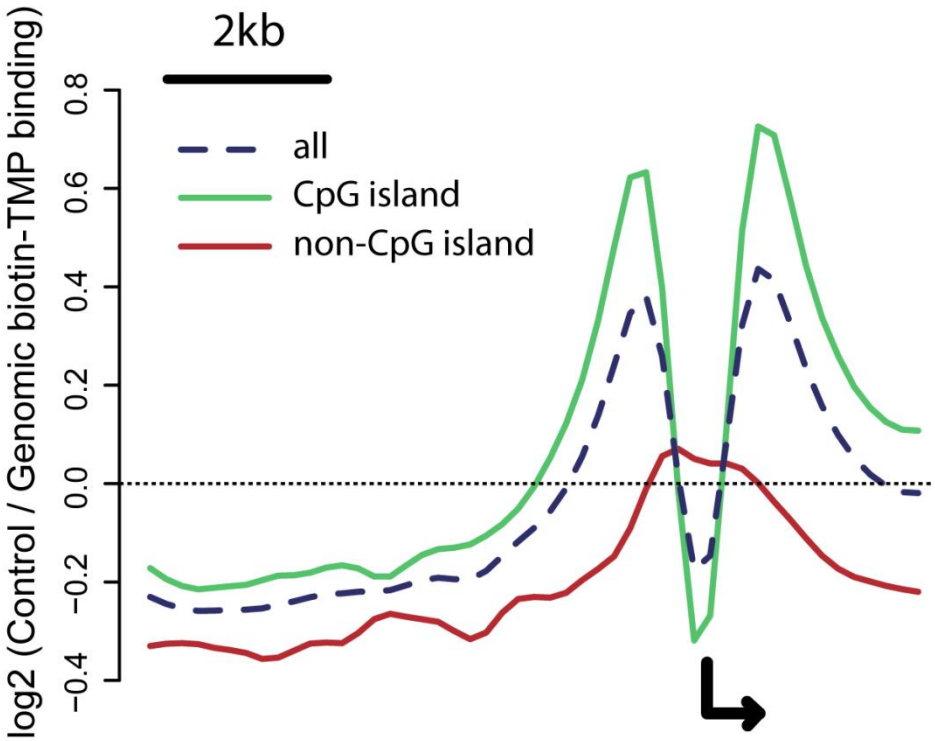


Figure 4.11 DNA supercoil distribution at CpG island and non-CpG island promoters. The median distribution of bTMP over 7 kb upstream to 3 kb downstream of human promoters genome wide (blue dashed line), CpG island promoters (green line) or non-CpG island promoters (red line). Corrected for bTMP sequence preference by subtracting the genomic IP sample. The position of the TSS is marked by the black arrow.

4.2.5 Non-CpG island promoter DNA supercoiling is extensively modified in expressed genes.

To identify the influence of gene expression on promoter structure genes were separated into highly expressed (top quartile) and non-expressed (bottom quartile) and the relative DNA supercoiling identified for CpG island and non-CpG island promoters (Figure 4.12). In both CpG island and non-CpG island gene promoters the relative distribution of DNA supercoiling is extensively modified compared to genomic DNA (dotted baseline through zero) for expressed and non-expressed genes. The distribution of DNA supercoiling around CpG island promoters is similar in expressed (4,286 TSSs) and non-expressed (4,186 TSSs) genes, displaying the characteristic under-wound structure 1 kb up-/down- stream of the TSS and relatively over-wound TSS. The magnitude of under-winding at expressed genes is elevated, supporting previous observations that the extent of DNA supercoiling is correlated with gene expression. The distribution of DNA supercoiling at non-CpG island promoters displays a more striking difference between expressed (963 TSSs) and non-expressed (4,947 TSSs) genes. In both expressed and non-expressed promoters there is a more under-wound TSS compared to the upstream region, however the magnitude and breadth of enrichment at expressed gene promoters is considerably more enriched than in CpG island promoters. Furthermore, non-expressed non-CpG island promoters have a relatively over-wound DNA structure, which may facilitate gene repression. This indicates that DNA supercoiling is strongly related to both expression and the underlying promoter sequence elements *in vivo*.

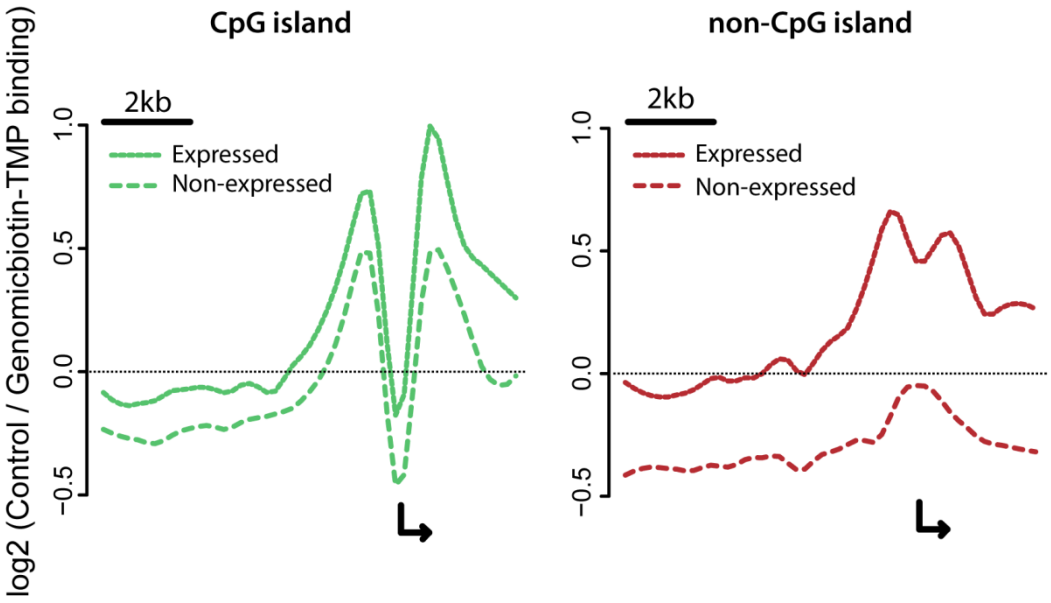


Figure 4.12 DNA supercoiling at expressed and non-expressed promoters. The median distribution of bTMP for expressed (expression array value > 155) and non-expressed (expression array value < 60) CpG island and non-CpG island promoters. The position of the TSS is marked by the black arrow.

4.2.6 CpG island and non-CpG island promoter DNA supercoiling is maintained by transcription

Previous studies have identified a relationship between transcription and DNA supercoiling at gene promoters. To establish if this relationship applies to the distinct structures observed at CpG island and non-CpG island promoters, the relative distribution of bTMP binding was compared between pull-down experiments performed on 'control', 'α-amanitin' and 'wash-out' samples (Figure 4.13). In each case the data was first normalised with genomic DNA bTMP pull-downs to give the *in vivo* DNA supercoil state. In samples treated with the RNA polymerase II inhibitor α-amanitin much of the DNA supercoil structure at gene promoters is lost, indicating that transcription maintains the structure of both CpG and non-CpG island promoters. The remaining DNA supercoil distribution following α-amanitin treatments may represent residual DNA supercoils that have not been released following transcription inhibition, or a structural property of CpG and non-CpG island promoters in the chromatin context which is absent from genomic DNA. The over-winding of DNA observed at CpG island promoters following α-amanitin treatment suggests that topoisomerases continue to remove DNA supercoils in the absence of transcription, in agreement with Naughton et al. (2013). The median distribution of non-CpG island promoters, on the other hand, becomes more under-wound in the presence of α-amanitin. This observation is consistent with previous work in the lab, although the mechanism through which DNA supercoils can be introduced independent of transcription are unknown.

Further confirmation of the relationship between transcription and promoter DNA supercoil structure comes from the structural recovery observed in the 'wash-out' sample. In this case cells were incubated for 3 hours in fresh media following treatment with α-amanitin. The DNA supercoil distribution of CpG island and non-CpG island promoters recovered, showing a much more similar distribution to the 'control' sample than the 'α-amanitin' sample. This is similar to the domain scale recovery seen by Naughton et al. (2013). Together this data shows that a peak of relatively under-wound DNA is formed at gene promoters with the onset of

transcription (‘wash-out’ sample) which is maintained under ‘steady state’ conditions (‘control’ samples).

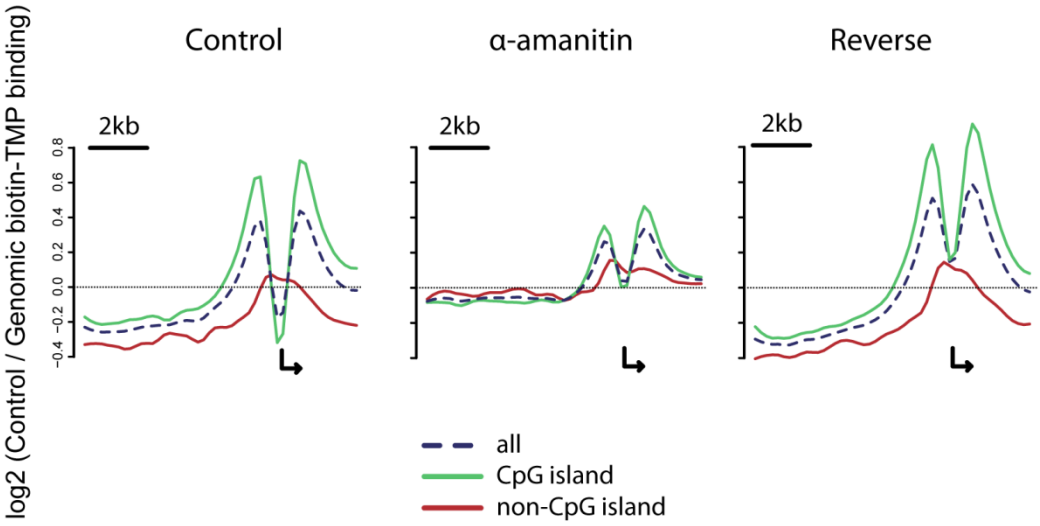


Figure 4.13 Transcription maintains promoter DNA structure. The DNA supercoil distribution at CpG island and non-CpG island promoters is remodelled by transcription inhibition and returns with the wash-out of the transcription inhibitor α-amanitin.

To identify if the DNA supercoil differences between the promoters of expressed and non-expressed genes are caused by transcription, the distribution of bTMP at expressed and non-expressed gene promoters was identified for the 'α-amanitin' samples. For both CpG and non-CpG island promoters the difference in DNA supercoil distribution between expressed and non-expressed genes is markedly reduced with α-amanitin (Figure 4.14). Therefore, the difference in DNA supercoiling between expressed and non-expressed genes is maintained by transcription. For CpG island promoters both expressed and non-expressed genes have a more over-wound DNA structure following transcription inhibition, indicating that CpG island promoter DNA supercoiling is generally maintained in an active/poised conformation. On the other hand, DNA supercoiling at active non-CpG island promoters is maintained by transcription in an under-wound conformation and at inactive non-CpG island promoters DNA is maintained in an over-wound conformation. The transcription dependent over-wound DNA structure of non-expressed non-CpG island promoters indicates that a repressive DNA supercoiling state is maintained by transcription at these genes and, in addition to the role of under-wound DNA in facilitating transcription, extends the function of gene regulation by DNA supercoiling to include repression.

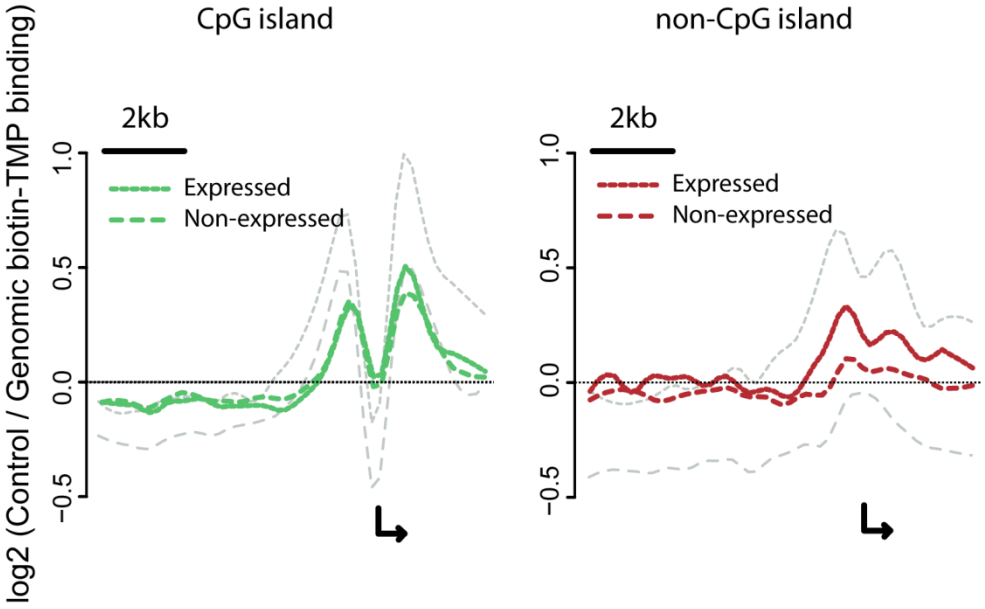


Figure 4.14 Expressed and non-expressed promoter DNA structure is maintained by transcription. DNA supercoiling distributions of α-amanitin samples for the expressed/non-expressed gene categories analysed in Figure 4.12 (shown in grey). Identifies a change in DNA supercoil structure at expressed and non-expressed genes with transcription inhibition.

4.2.7 Promoter DNA supercoiling identifies expression differences independent of the underlying sequence composition

DNA supercoiling at CpG and non-CpG island promoters is distinct (Section 4.2.4). To further investigate promoter DNA structure, CpG island and non-CpG island promoters were each classified based on DNA supercoil distribution using a k-means cluster analysis with 3 clusters. By separating promoters based on DNA supercoil distribution, the relative influence of sequence and expression on promoter supercoiling can be determined.

4.2.7.1 CpG island promoters

The classification of CpG island promoters by DNA supercoil distribution identifies promoter structures that vary at both the TSS and across the locus, independent of one another (Figure 4.15a). For example 'class 1' and 'class 3' promoters have a similar distribution of DNA supercoiling at the TSS but very different distributions from 1 kb up-/down- stream. 'Class 1' and 'class 2' on the other hand have similar distributions of DNA supercoiling 5 kb upstream but very different distributions at the TSS. To identify how sequence and gene expression relate to these promoter classes, nucleotide composition and expression profiles were determined for each.

The identification of a distinct over-wound DNA supercoil distribution at CpG islands (Figure 4.15a) indicates a relationship between sequence and DNA supercoil distribution. To determine whether the pattern of DNA supercoiling is purely a reflection of sequence distribution, or represents a distinct structural feature at gene promoters, the median GC% distribution was plotted for each CpG island promoter class (Figure 4.15b). The level of over-wound DNA supercoiling at the CpG island is related to GC%, with more GC rich CpG islands being more over-wound. However, the peaks of under-wound DNA identified in 'class 1' CpG island promoters cannot be accounted for by gross differences in sequence, as the GC% of all three CpG island promoter classes is highly similar in the region 1 kb up-/down-stream of the TSS. Furthermore, 'class 1' and 'class 3' promoters show the largest

difference in DNA structure but have almost identical GC% distributions. This confirms that DNA sequence and DNA supercoiling structure are distinct phenomenon in the *in vivo* chromatin context. On the other hand, 'class 2' promoters have a distinct sequence distribution and DNA structure, and in this case the more intense peak of GC% at the TSS of 'class 2' genes may account for the more over-wound DNA structure observed by bTMP binding. Together, this data indicates that promoter DNA supercoil distribution is related to sequence, but that sequence alone cannot explain the intensity of DNA supercoiling in 'class 1' or 'class 3' promoters. This supports supercoiling as an independent structural feature of human DNA.

Previous analysis identified a relationship between expression and DNA supercoil distribution, with under-wound promoters being more highly expressed (Section 4.2.5). To identify the relationship between expression level and DNA supercoil distribution in the three promoter classes, the relative expression of the genes in each class was determined from expression array data (Figure 4.15c). 'Class 1' promoters are associated with high expression level, while 'class 2' and 'class 3' promoters have lower expression. The major feature of 'class 1' CpG island promoters is the intense under-wound DNA peaks 1 kb up-/down- stream of the TSS, which are absent in both 'class 2' and 'class 3' promoters. Therefore, the DNA supercoil distribution of highly expressed gene promoters ('class 1') is distinct from lowly expressed gene promoters ('class 2' and 'class 3'), but expression cannot be used to determine difference in DNA supercoiling between 'class 2' and 'class 3' promoters.

This analysis determined that differences in DNA supercoil distribution between 'class 1' and 'class 3' promoters are not due to DNA sequence distribution, but can be explained by differences in gene expression. Conversely, the difference between 'class 2' and 'class 3' promoters cannot be explained due to differences in expression, but can be explained by differences in DNA sequence. Together, this identifies that DNA sequence and gene expression are both important for the distribution of DNA supercoiling at gene promoters.

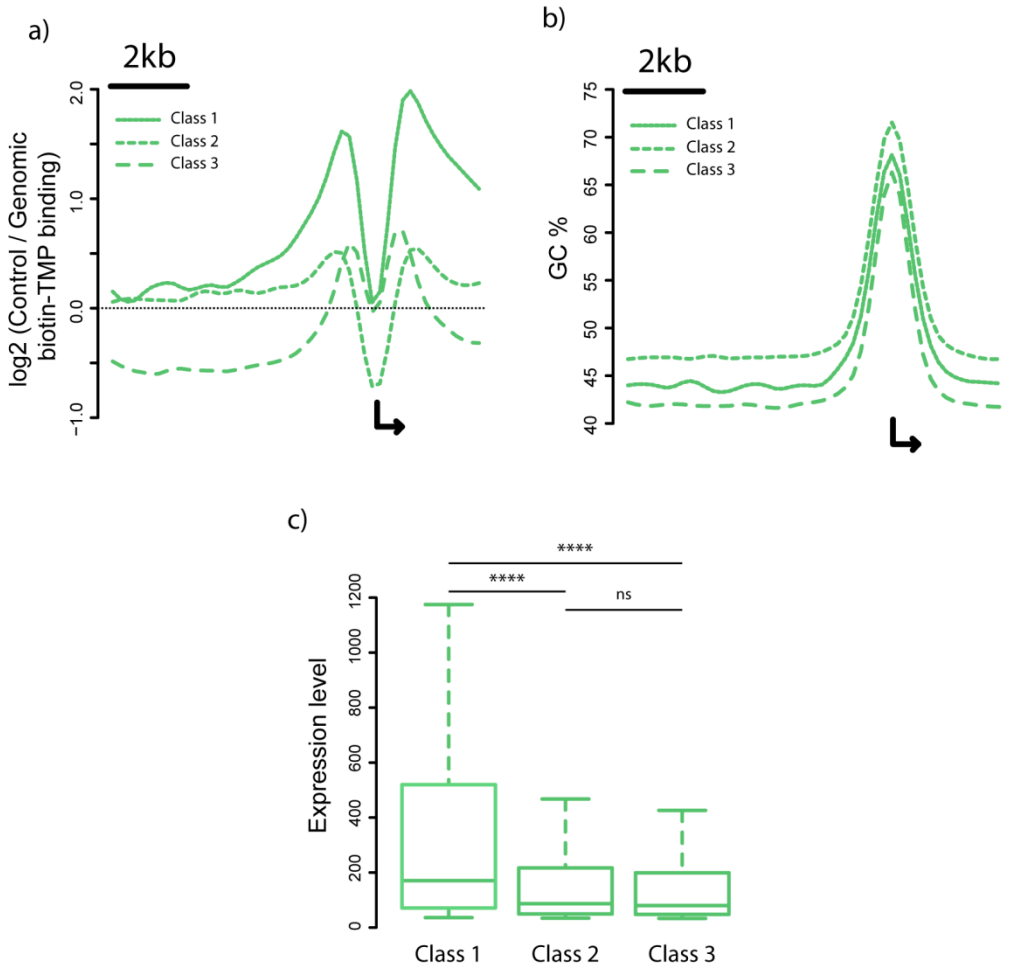


Figure 4.15 CpG island promoter classification. a) K-means clustering of CpG island promoters into 3 classes based on bTMP binding distribution. b) GC percentage sequence distribution for the three CpG island promoter classes. c) Boxplot of gene expression levels in RPE1 cells for the three CpG island promoter classes.

4.2.7.2 Non-CpG island promoters

A similar classification was performed for non-CpG island promoters to identify the relationship between DNA structural features, sequence and expression. This clustering analysis identified promoter structures with a very different distribution to those observed for CpG island promoters (Figure 4.16a). The strongest determinant of class membership is the overall level of bTMP bound across the 10 kb window, therefore separating promoters based on local DNA structure rather than the specific structure at the transcription start site. All three classes of promoter have a relatively under-wound peak at the TSS, although this is more prominent in 'class 1' and 'class 3' promoters. To determine whether the DNA supercoil distribution at these clusters is determined by sequence, the median distribution of GC% was plotted for each class of promoters (Figure 4.16b). 'Class 1' and 'class 2' promoters have highly similar sequence distributions whereas 'class 3' promoters show the same pattern of sequence distribution, but with a much lower GC% magnitude. Therefore, non-CpG island promoters also show differences in DNA supercoil distribution that cannot be attributed to sequence, further supporting supercoiling as an independent structural feature of DNA.

To identify the relationship between non-CpG island promoter DNA structure and gene expression, the relative expression level of genes within the promoter classes was determined (Figure 4.16c). Similar to previous observations (Section 4.2.5), an over-wound DNA structure is associated with gene repression and an under-wound structure is associated with gene expression. The expression difference between each of the non-CpG island promoter classes is highly significant, indicating that expression level can be used to differentiate between these promoter classes.

Together this data shows that for non-CpG island promoters, gene expression more clearly differentiates promoter class than sequence. This is particularly clear for 'class 1' and 'class 2' promoters which have very similar sequence distributions, but distinct expression levels. For 'class 3' promoters the role of expression and sequence cannot be separated, as these promoters have a distinct sequence distribution and gene expression level. Therefore, sequence is likely to have some

influence on DNA supercoiling, but expression is a more important determinant of promoter DNA supercoil class at non-CpG island promoters.

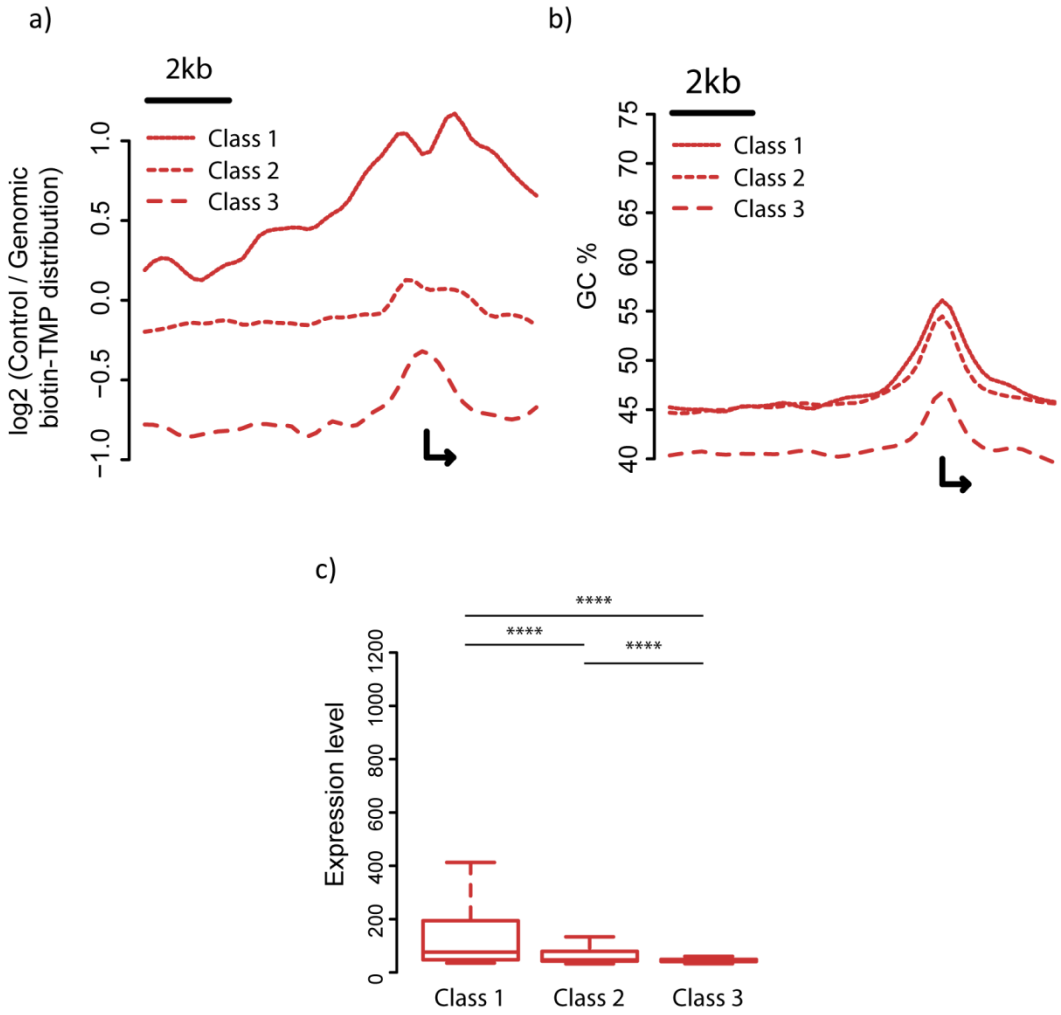


Figure 4.16 Non-CpG island promoter classification. a) K-means clustering of non-CpG island promoters into 3 classes based on bTMP distribution. b) Boxplot of gene expression levels in the three non-CpG island promoter classes. c) GC percentage sequence distribution for the three non-CpG island promoter classes.

Together, the distribution and classification of DNA supercoiling at CpG island and non-CpG island promoters identifies a number of common themes. In each case there is a class with a different sequence distribution which has a distinct DNA structure and a low expression level. However, in the case of CpG island promoters this class has a higher GC content whereas in the case of non-CpG islands this class has a lower GC content. Together this suggests that it is not the sequence per se that is repressing gene expression, but the DNA structure itself. In which case, this identifies a GC and an AT associated repressive DNA structure. The pattern of DNA structural distribution between the remaining two classes of CpG and non-CpG island promoters identify an average difference in bTMP cross-linking of around 2 fold between the expressed and non-expressed class, independent of sequence. This is consistent with the observations of Kouzine et al. (2013) and indicates that CpG island and non-CpG island promoters have an expressed DNA supercoil distribution that can be detected through an unsupervised classification of supercoil distribution.

4.2.8 Generally expressed genes maintain an 'active' DNA supercoil distribution independent of expression

Genes can be categorised into those that are expressed in most human cell types and those that are not. The features that differentiate generally expressed and generally repressed genes are not well characterised, although genomic structure and epigenetic modifications have been implicated (Eisenberg and Levanon, 2003; She et al., 2009). To identify if generally expressed/repressed genes have distinct DNA supercoil structure, the distribution of bTMP was identified for these categories based on the expression data of 43 normal human tissues (Chang et al., 2011).

Generally expressed genes are more under-wound with a CpG island like distribution whereas generally repressed genes are more over-wound with a less pronounced dip at the TSS (Figure 4.17). Expression array data for our cell line supports the gene classification, with the generally expressed class being much more highly expressed than the generally repressed class. To confirm that the underlying structure of the generally expressed genes is mostly CpG island promoters and the generally

repressed is mostly non-CpG island promoters, as indicated by the distributions, the relative CpG/non-CpG island level was established. For generally expressed promoters 84% have a CpG island compared to 37% of generally repressed promoters. Therefore, although CpG islands are associated with generally expressed genes, the presence of CpG islands at generally repressed genes indicates that this sequence parameter alone is not sufficient to identify if a gene has a general expression in multiple cell lines.

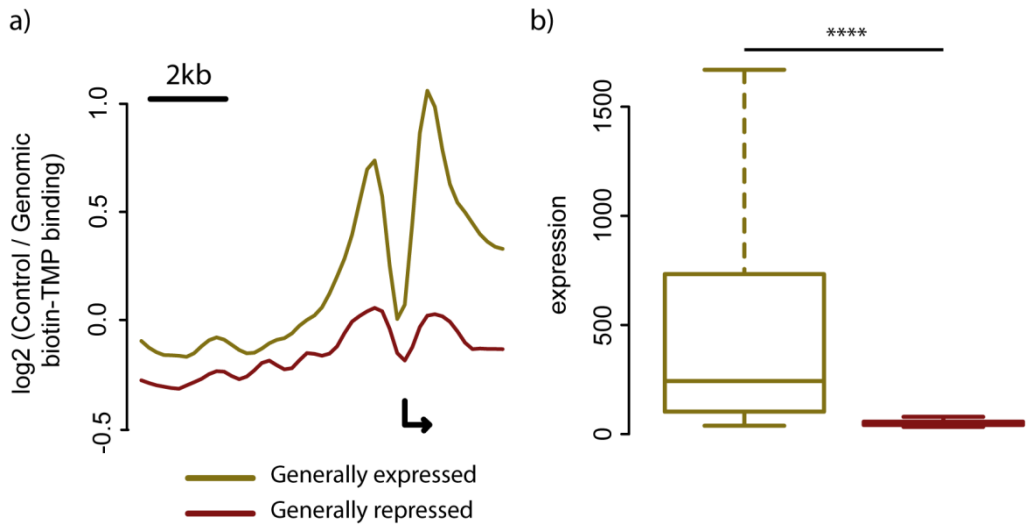


Figure 4.17 DNA supercoil distribution at generally expressed and repressed genes. a) DNA supercoil distribution around the TSS of generally expressed (2281 TSSs) and non-expressed (2961 TSSs) genes. b) Relative expression of generally expressed and generally repressed genes. **** represents a p-value<0.00005.

To identify if a specific DNA supercoil pattern for generally expressed and generally repressed promoters exists, the distribution of bTMP at these gene promoters was investigated more thoroughly. Previous analyses have shown that CpG island and non-CpG island promoters have distinct structures (Section 4.2.5), therefore further analysis of generally expressed and generally repressed genes was performed separately on these promoter classes. For generally expressed genes CpG island (2492 TSSs) and non-CpG island (459 TSSs) promoters have a similar magnitude of bTMP binding, but a different distribution, indicating similar levels of under-wound supercoiling distributed in a different promoter structure (Figure 4.18a). The distribution of generally repressed CpG island promoters (846 TSSs) maintains an under-wound DNA structure around the TSS, although the magnitude is reduced compared to that of 'generally expressed' promoters. On the other hand, the distribution of 'generally repressed' non-CpG island promoters (1435 TSSs) is more over-wound. Together the CpG and non-CpG island data supports previous observations that expressed genes are more under-wound (Section 4.2.5; Kouzine et al., 2013; Naughton et al., 2013a).

To establish the relationship between the structure of generally expressed genes and the structure of genes expressed in RPE1 cells, a comparison was made between genes that are generally expressed but are not expressed in RPE1 cells and genes that are generally repressed but are expressed in RPE1 cells (Figure 4.18b). Surprisingly, both CpG and non-CpG island promoters for generally expressed genes maintain their structure independent of whether they are expressed in RPE1 cells. Both the magnitude of bTMP intercalation and its distribution around the TSS are highly similar, supporting the existence of a generally expressed DNA promoter structure. On the other hand, the generally repressed promoters have very different structures between those that are expressed and non-expressed in RPE1 cells, with expressed genes having a more under-wound structure. Sequence distribution analysis identifies that the difference observed between expressed and non-expressed genes in the four categories ('generally expressed' CpG/nonCpG, 'generally repressed' CpG/non-CpG) cannot be accounted for by differences in the GC% distribution at the promoter. Therefore, there is a DNA supercoiling component of generally expressed gene structure, which cannot be accounted for by sequence or gene

expression. This suggests that genes which are expressed in most cell types are maintained in a DNA supercoiling conformation associated with gene activity, whereas genes which are generally repressed in most cell types have a generally repressive structure which is remodelled to an active DNA supercoil distribution with expression.

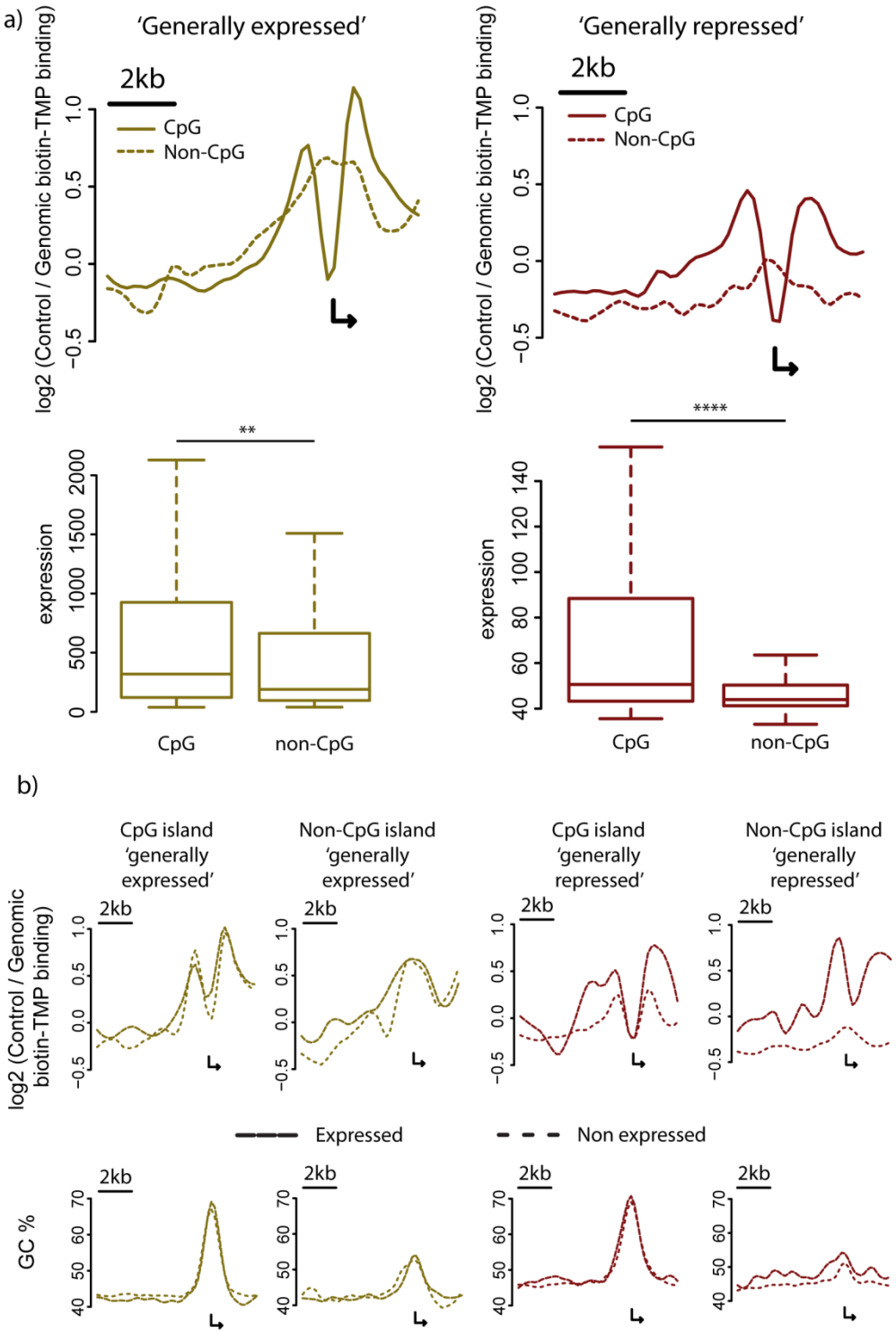

a)
'Generally expressed'
'Generally repressed'
2kb
CpG
Non-CpG
log2 (Control / Genomic biotin-TMP binding)
expression
**
****
CpG
non-CpG
b)
CpG island 'generally expressed'
Non-CpG island 'generally expressed'
CpG island 'generally repressed'
Non-CpG island 'generally repressed'
log2 (Control / Genomic biotin-TMP binding)
Expressed
Non expressed
GC %

Figure 4.18 Generally expressed genes have a more under-wound DNA structure at gene promoters, independent of gene expression. a) DNA supercoiling distribution and expression of generally expressed and generally repressed genes for CpG and non-CpG island promoters. The expression levels of generally expressed (gold) and generally repressed (red) genes are on different scales to show the difference between CpG and non-CpG islands promoters. Student's t-test performed between datasets with a p-values of **<0.005 and ****<0.00005. Generally expressed CpG – 2492 TSSs, generally expressed non-CpG – 459 TSSs, generally expressed non-CpG – 846 TSSs and generally repressed non-CpG – 1435 TSSs. b) DNA supercoiling and sequence distribution at generally expressed and generally repressed promoters classified on actual expression in our RPE1 cells. Expressed promoters have a value greater than the third quartile of the expression array dataset, whereas non-expressed genes have an expression less than the first quartile. Loess smoothing applied. Expressed 'generally expressed' CpG island – 1728 TSSs, non-expressed 'generally expressed' CpG island – 222 TSSs, expressed 'generally expressed' non-CpG island – 273 TSSs, non-expressed 'generally expressed' non-CpG island – 62 TSSs, expressed 'generally repressed' CpG island – 125 TSSs, non-expressed 'generally repressed' CpG island – 524 TSSs, expressed 'generally repressed' non-CpG island – 61 TSSs and repressed 'generally repressed' non-CpG island – 1199 TSSs.

4.2.9 Transcription factor, enhancer and insulator binding sites have distinct DNA supercoiling profiles.

DNA supercoiling has been shown to regulate transcription through changes in DNA structure and through the recruitment of DNA structure specific transcription factors. To identify if DNA binding proteins have distinct DNA supercoiling profiles, the distribution of bTMP binding was determined at transcription factors, enhancer and insulator binding sites. Using ChIP-seq data available from the ENCODE consortium for the A549 epithelial cell lines the position of actual protein binding sites *in vivo*, as opposed to consensus binding sequences, was determined for three transcription factors, two enhancer proteins and an insulator protein. The distribution of DNA supercoils 2.5 kb either side of the protein binding site was then determined from our RPE1 cell line (Figure 4.19). The transcription factors each show an enrichment for under-wound DNA around the binding site, with a more over-wound structure at the protein binding site itself. This distribution is most similar to that of active genes, although there is some difference between the three transcription factors. ELF1 in particular shows a more under-wound structure across the region analysed. The distribution of DNA supercoiling at enhancer proteins is less consistent, with p300 showing some similarities with transcription factors whereas CEBP1 has a very different distribution. The p300 distribution shows a peak of under-wound DNA supercoiling 1 kb from the protein binding site and a less intense trough at the binding site than observed for transcription factors. On the other hand, CEBP1 has a gradual low level enrichment of under-wound DNA associated with the protein binding site. Finally, there is almost no peak of under-wound DNA supercoiling at promoter associated CTCF binding sites. The slight peak of under-wound DNA 1 kb from the binding site and slight over-winding at the binding site are somewhat similar to the distribution seen at transcription factors and p300, but with a very low magnitude. Together this data identifies that different DNA binding proteins have a different distribution of DNA supercoiling around their biding sites in the context of chromatin.

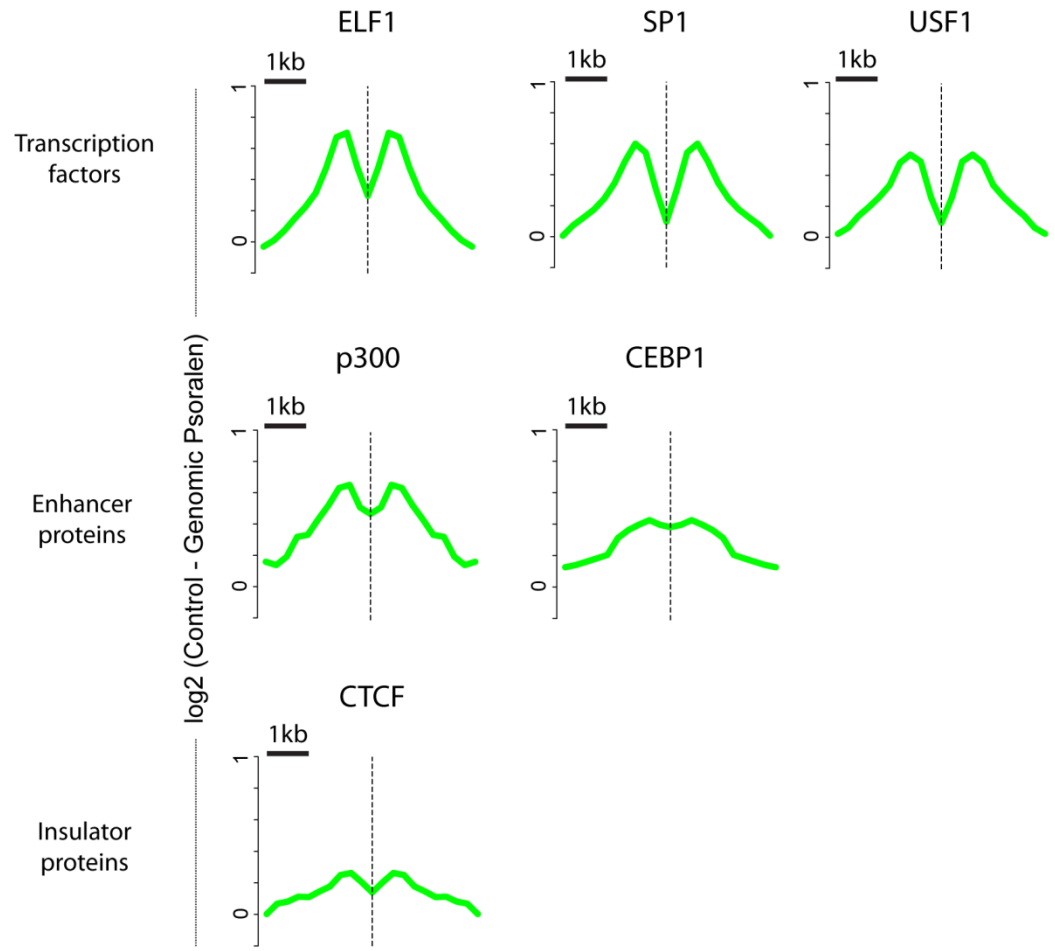


Figure 4.19 DNA supercoiling around protein binding sites. Protein binding sites determined from ChIP-seq data in A549 epithelial cells from the ENCODE project. Distribution identifies distribution up to 2.5 kb from the TSS.

4.3 Discussion

The current study provides the first genome-wide analysis of the distribution of DNA supercoiling at human gene promoters. The increased resolution of a promoter scale analysis of DNA supercoiling, when compared to the domain scale analysis of Naughton et al. (2013), necessitated a more detailed characterisation of the sequence preference of the bTMP molecule. Photo-crosslinking bTMP with DNA oligonucleotides of defined sequence (e.g. poly-d(AT) compared to poly-d(GC)), and on genomic DNA, identified interactions with the thymine only. This supports previous studies that identify that TMP has a preference for thymine (Esposito et al., 1988; Kanne et al., 1982; Song and Ou, 1980). These experiments cannot rule out an interaction between bTMP and cytosine, which may occur infrequently and below the threshold of detection. Supporting this, the one study that identifies TMP bound cytosine performed the photo-crosslinking reaction several times with a very high concentration of TMP prior to analysis by mass spectrometry (Kanne et al., 1982). Consequently, unless the biotin and charged linker moieties altered the affinity of TMP for cytosine, it is unlikely that under my experimental conditions I would be able to detect bTMP bound cytosine. The thymine preference of bTMP is not unusual for a DNA binding molecule. For example, the ubiquitous cross-linking reagent formaldehyde only binds to guanine nucleotides (Lu et al., 2010) and the topoisomerase II poison etoposide preferentially stabilises cleavage complexes at cytosine nucleotides (Wu et al., 2011). However, it is important to take any nucleotide bias into account for subsequent analyses.

To determine whether this thymine preference alters the pull-down efficiency of genomic DNA following bTMP photo-crosslinking, which could bias subsequent analysis, bioinformatic analysis was performed on bTMP bound genomic DNA pull-down experiments. A subtle negative correlation was observed between GC% and bTMP, indicating a subtle bias, although a strong depletion was not observed even at high GC% and structural factors in addition to thymine preference could account for lower bTMP binding (e.g. the formation of Z form DNA (Thamann et al., 1981)). In order to identify structural features which may influence bTMP binding, an

experiment was performed which compared the relative photo-crosslinking of bTMP to oligonucleotides with defined A, B or AB form DNA structures. Under these experimental conditions no difference attributable to DNA structure was identified, although the limited availability of oligonucleotides with known helical structures prevented a more thorough analysis of the relationship between bTMP binding and DNA helical structure. However, it has been shown that psoralen can have a sequence preference that is more complex than the proportion of GC (Esposito et al., 1988), therefore to account for this unknown bias bTMP pull-down experiments (e.g. control, α-amanitin and reverse) were corrected at a microarray probe level for bTMP enrichment in genomic DNA.

The absence of a strong sequence bias in the analysis of bTMP pull-down microarray data was surprising, given the bias observed in earlier experiments, but has been seen previously in our lab. It is possible that at the scale of 300bp fragments the relationship between bTMP and sequence is distinct from that of bTMP and single nucleotides (such as the HPLC MS experiment in Section 4.2.1.2). Furthermore, the chemical structure of cytosine should permit psoralen photo-crosslinking, and has been shown to do so under high psoralen conditions (Kanne et al., 1982). The HPLC-MS experiments that aimed to identify bTMP sequence preference in heterogeneous DNA sequences may not have been sensitive enough to pick up a smaller population of bTMP bound cytosine, which could have a significant influence on the enrichment of high GC fragments.

To test more thoroughly the relationship between bTMP binding and DNA sequence within genomic DNA, I am currently performing a number of additional bioinformatic analyses and laboratory experiments. Firstly, a new dataset has become available in which psoralen binding has been established genome-wide in *Drosophila* (Teves and Henikoff, 2014). Through a detailed analysis of this data I can validate the observed relationship between genomic DNA and psoralen binding. In addition, I am performing an experiment in which bTMP-bound DNA is fractionated based on sequence through a CsCl gradient (Scott et al., 2005) and the level of bTMP quantified by dot-blot. This experiment is analogous to the bioinformatic analysis performed in Section 4.2.2.2. Performing this gradient

analysis on bTMP bound naked DNA and on DNA extracted from RPE1 cells in the presence and absence of the nicking agent bleomycin will give an additional, albeit crude, indication of how chromatin influences bTMP-DNA binding at sequences with different GC compositions. Together, these analyses will be used to validate the observations and conclusions presented in this chapter.

My genome-wide analysis of promoter DNA supercoiling supports the transcription dependent under-wound DNA structure identified in Kouzine et al. (2013) and Naughton et al. (2013a). One clear difference between the distribution observed in my genome-wide study, when compared with previous studies, is the presence of a more over-wound structure at the transcription start site compared to the more under-wound regions ~1 kb up- and down-stream. This structure may not have been picked up in previous studies for several reasons including the application of smoothing algorithms, the investigation of a wider domain around the TSS and the relatively small number of transcription start sites assayed. In this study there is no smoothing algorithm applied to the genome-wide analysis of promoter DNA supercoiling, and as such it is a true representation of the average supercoil state.

To characterise further the distribution of DNA supercoiling at the start of genes, promoters were classified based on known characteristics. A well characterised property of most human promoters is the presence of a CpG island, which is believed to have an important role in gene regulation (Ehrlich et al., 1982). Ranking promoters on their probability of containing a CpG island (Irizarry et al., 2009; Wu et al., 2010) identifies two distinct promoter DNA supercoil distributions and it is immediately apparent that the over-wound DNA structure observed at the TSS is associated with CpG islands. At CpG island promoters the relatively over-wound TSS may repress gene transcription and it may be that CpG islands act as a general transcription repressor which must be overcome prior to gene expression, perhaps through the activity of DNA helicases (Singleton et al., 2007). As genes with CpG islands are generally highly expressed, it may be that the over-wound CpG island DNA prevents leaky expression in a chromatin environment that promotes transcription. On the other hand, non-CpG island gene are generally tissue specific and may be regulated primarily by other means, such as the concentration of

appropriate transcription factors. As the DNA supercoil distribution of CpG and non-CpG island promoters was so clearly distinct, all subsequent analysis was performed separately for these classes of promoter.

A second known characteristic of gene promoters in RPE1 cells is the gene expression level, as identified by expression array analysis. Previous studies have identified for a few hundred gene promoters that expressed genes have a more under-wound DNA structure (Kouzine et al., 2013; Naughton et al., 2013a). To identify the influence of expression on CpG island and non-CpG island promoter structure, the distribution of DNA supercoiling was determined for the promoters of expressed and non-expressed genes. For CpG island promoters the expressed and non-expressed genes each show a highly similar distribution, but with a relatively subtle enrichment of under-wound DNA in the expressed genes over non-expressed. This enrichment of under-wound DNA occurs across the 10 kb measured, indicating that the CpG islands of expressed genes are generally more under-wound than their non-expressed counterparts. This could influence gene expression by relieving some of the repression that may be associated with the relatively over-wound DNA structure of CpG islands. There is a much more extensive rearrangement in DNA structure in non-CpG island promoters between non-expressed and expressed genes, with a greater enrichment for under-wound DNA in the expressed genes. The DNA supercoil distribution of expressed genes resembles that of CpG island promoters, suggesting that active gene promoters generally have an under-wound DNA supercoil structure that peaks in a region ~1 kb up-stream to ~1 kb down-stream. In addition, the over-wound DNA supercoil structure of non-expressed non-CpG island promoters could represent a novel form of transcription regulation in eukaryotic genome through the repression of strand separation. Several reports have discussed the positive regulatory potential of more under-wound DNA at gene promoters (Dunaway and Ostrander, 1993; Hirose and Suzuki, 1988; Kouzine et al., 2004, 2008, 2013; Naughton et al., 2013a, 2013b; Tabuchi and Hirose, 1988), but no study has explicitly discussed the potential of relatively over-wound DNA in the repression of gene expression. To identify if the relationship between gene expression and DNA supercoil distribution is a consequence of transcription, the distribution of bTMP was identified in the same gene categories following transcription inhibition.

Under these conditions, the under-wound DNA of CpG island/expressed non-CpG island promoters and the over-wound DNA of non-expressed CpG island promoters is lost. Together, this supports a model that extends the role of transcription dependent DNA supercoiling from under-wound DNA facilitating expression to include over-wound DNA maintaining repression.

To determine the attributes of DNA supercoiling at CpG island and non-CpG island promoters independent of known parameters, promoters were clustered based purely on the distribution of DNA supercoils by kmeans analysis. For both CpG island and non-CpG island promoters it is clear that sequence has an important influence on DNA supercoil distribution, but does not determine it. For example, in both CpG island and non-CpG island promoters there are promoter classes with identical sequence distributions but distinct DNA supercoil distributions, and in both cases the more under-wound class has a higher expression level. Furthermore, in both CpG and non-CpG island promoters there is one class with a substantially different sequence distribution, and in both cases this sequence is associated with lower expression. Surprisingly, the 'repressive' sequence distribution in CpG island promoters is GC rich whereas in non-CpG island promoters it is GC poor, indicating that GC and AT associated repressive DNA structures occur *in vivo*. The GC associated repressive structure may be accounted for by the more over-wound DNA at the TSS, which in my model of transcriptional repression by CpG islands would require additional rearrangement for strand separation at the TSS. In the case of AT-rich repressive structures, the higher flexibility of AT-rich sequences may be more amenable to over-wound DNA supercoils in the generation of a repressive DNA supercoil environment. Together, this data clearly demonstrates that differences in promoter DNA supercoiling are determined by the combined influence of sequence and gene expression. This suggests a model whereby gene sequence can produce a permissive environment for the storage of under- or over- wound DNA supercoils that can facilitate gene expression or repression *in vivo*.

To test whether permissive or repressive DNA supercoil distributions are associated with genes that are generally expressed or repressed across a wide range of tissue types, a comparison was made between the distribution of DNA supercoiling at

genes which are generally expressed/repressed and expressed in our RPE1 cells and genes which are generally expressed/repressed and not expressed in our RPE1 cells. This analysis determined that generally expressed genes have a permissive under-wound DNA structure independent of expression and generally repressed genes are only under-wound when expressed in our cell type. This supports DNA supercoiling as an additional factor in establishing the promoter structure of generally expressed genes. In addition, the distribution of DNA supercoiling at generally repressed genes further supports a model whereby transcription remodels DNA supercoiling to a more under-wound state that is permissive for subsequent transcription from the same promoter.

In addition to the direct influence of DNA supercoiling on transcription from gene promoters, through an increased efficiency of initiation and elongation (Ma et al., 2013a; Tabuchi and Hirose, 1988), it is likely that supercoil dependent changes in DNA structure influence transcription factor binding. For example, the melting of the FUSE element upstream of Myc by under-wound DNA supercoiling results in the binding of FBP (Kouzine et al., 2008). To identify if distinct DNA supercoil distributions can be seen at promoter associated DNA binding proteins, an analysis of bTMP was carried out on regions surrounding transcription factor binding sites and structural proteins. This analysis indicates that transcription factors occupy sites with distinct DNA supercoiling profiles from other DNA binding proteins, such as the structural protein CTCF. Furthermore, there is a subtle difference in DNA supercoil distribution between the transcription factors ELF1, SP1 and USF1 which may be indicative of different supercoil preferences. To determine with precision how DNA sequence and structure work together with DNA supercoiling to form a preferential substrate for DNA binding proteins, future work must determine the distribution of DNA supercoiling at the scale of DNA binding protein sequence motifs (i.e. through deep sequencing).

The aim of this study was to characterise the distribution of DNA supercoiling at human gene promoters. The most striking observation is the distinct structures of CpG and non-CpG island promoters and in both cases a model of gene regulation through DNA supercoiling is proposed. In addition, an analysis of DNA supercoil

distribution around DNA binding proteins tentatively supports the hypothesis that different DNA binding proteins bind different DNA structures *in vivo* and necessitates the investigation of more proteins by ChIP-seq and a higher resolution mapping of bTMP using next generation sequencing. Together, this data confirms a relationship between DNA supercoiling, sequence and gene expression, supporting DNA supercoil distribution as an important genetic regulator at the promoters of human genes.

5 DNA supercoiling at common fragile sites

5.1 Introduction

Common fragile sites (CFSs) are conserved regions of human chromosomes that form breaks, constrictions, gaps and rearrangements following partial DNA replication inhibition (Section 1.3.1). The mechanism of fragility at these sites is unknown, although several molecular characteristics implicate DNA supercoiling. Supporting this, I have identified regions of stable over-wound DNA supercoiling and topoisomerase depletion at the FRA3B and FRA16D CFSs (Section 3.2.7). This indicates that these CFSs have a reduced ability to maintain DNA supercoils under normal culture conditions, which may be accentuated by replicative stress. However, changes in DNA supercoiling have not been measured under conditions of replicative stress and the relationship with CFSs remains theoretical. To determine whether DNA supercoiling may contribute directly to the instability of CFSs, a bTMP pull-down approach was used to measure changes in DNA supercoils at the FRA3B and FRA16D loci following replication inhibition.

5.1.1 Molecular properties of CFSs

The distribution and molecular basis of CFSs are generally determined following the partial inhibition of DNA replication with the DNA polymerase inhibitor aphidicolin (Lukusa and Fryns, 2008). Under these conditions, metaphase chromosomes exhibit aberrations clearly visible by light microscopy (expressed CFSs). Additionally, in interphase aphidicolin treatment causes DNA damage, which is thought to contribute to CFS activity and can be detected by an increase in the number of γH2AX foci (Schwartz et al., 2005). Why partial replication inhibition reproducibly damages DNA and activates fragility at defined loci is largely unknown, although certain properties of CFSs have been determined.

To map the location of expressed CFSs in metaphase, the DNA is stained with either giemsa or DAPI to give a precise banding pattern to the chromosomes, based on an affinity for AT-rich DNA (Gosden, 1994). The chromosome karyotype is then used to determine where on a chromosome a particular CFS lies, for example FRA3B is found within band 3p14.2 and FRA16D is found within 16q23.2 (Figure 5.1). Using this technique to determine the distribution of CFSs in human lymphoblastoid, fibroblast, epithelial and erythroid cells, it has become apparent that the complement of expressed CFSs is cell type dependent (Le Tallec et al., 2011, 2013). For example, in lymphoblastoid cells FRA3B and FRA16D are the most expressed CFSs whereas in fibroblasts it is 3q13.3 and 1p31.1.

To identify the cytological position of CFSs (Figure 5.1 karyotype) on the DNA sequence (Figure 5.1 genomic position) FISH based approaches have been used. In these studies fluorescent DNA probes with sequences tiling the band of interest are hybridised to aphidicolin treated metaphase chromosomes and investigated for their proximity to the CFS break point. This enabled the identification of CFSs as large unstable domains covering as much as 4 Mb, such as at FRA3B (Becker et al., 2002). To determine the sequence properties of these CFS domains, the boundaries determined by FISH have been mapped onto the human genome. In general, CFSs are AT-rich with a highly flexible DNA structure (Lukusa and Fryns, 2008) which can melt easily and form alternative DNA structures in the presence of DNA supercoiling (Bacolla et al., 1997; Burrow et al., 2010; Zlotorynski et al., 2003). CFSs are generally associated with long genes (> 300 kb), for example FHIT at FRA3B and WWOX at FRA16D, and are enriched for AT-rich DNA (Helmrich et al., 2006; Le Tallec et al., 2013). There is some controversy as to whether transcription at these long genes increases chromosome fragility through the collision of polymerases and replication fork arrest, with evidence both for (Helmrich et al., 2006) and against (Le Tallec et al., 2013). However, the mapping of many CFSs to regions close to, rather than within, large genes suggests that transcription per se does not set the borders of CFSs (Le Tallec et al., 2013). The identification of large genes and DNA structural properties associated with CFS formations suggests that sequence is an important factor. However, the cell-type specificity of CFSs indicates that sequence is not the sole determining factor of CFS expression.

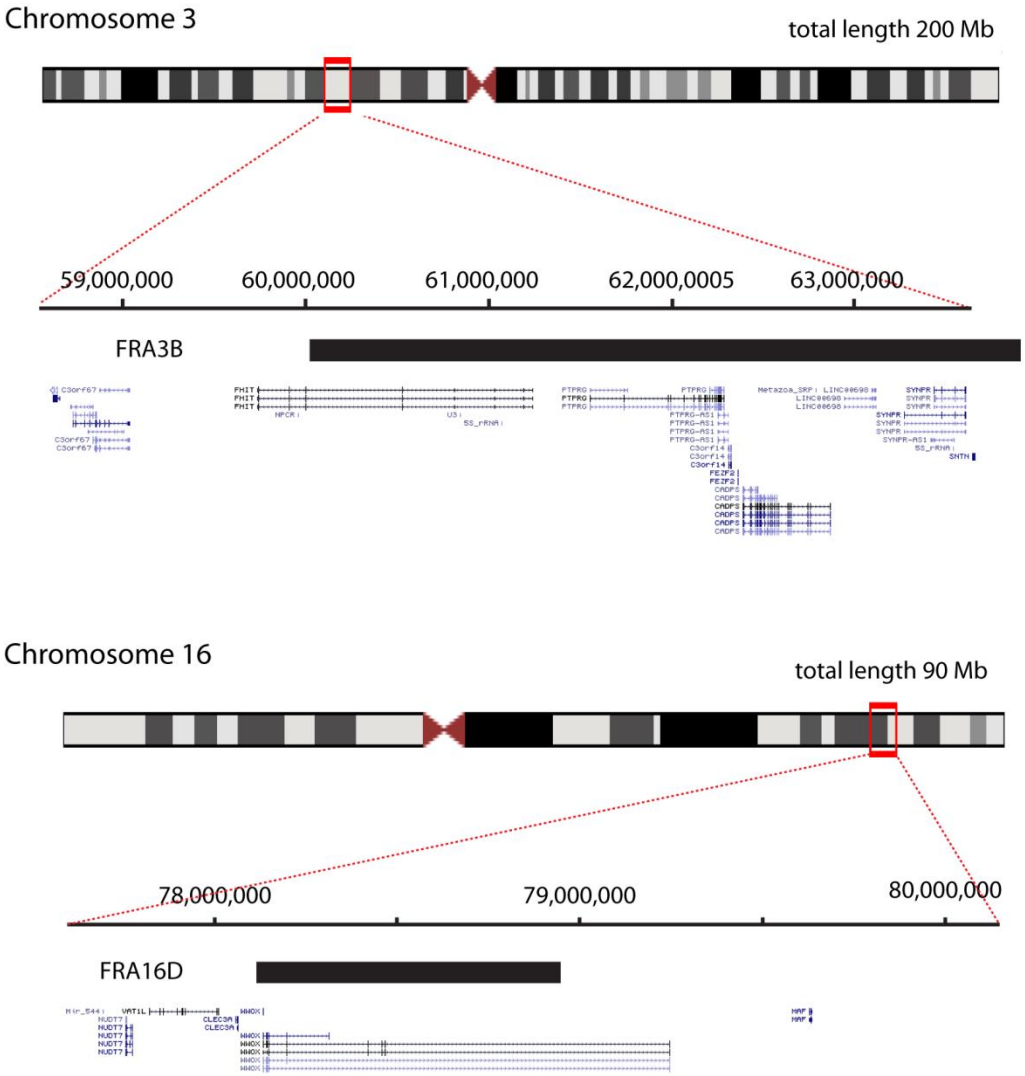


Figure 5.1 The position of FRA3B and FRA16D common fragile sites. The karyotype for human chromosome 3 and 16 with the position of CFSs FRA3B (3p14.2) and FRA16D (16q23,2) as identified in Becker et al. (2002) and Letessier et al. (2011). Chromosome length is given at the top right of each banded chromosome.

To determine how changes in replication timing following aphidicolin treatment relate to the distribution of CFSs in a particular cell type, the origin and progression of DNA replication was determined at several loci in lymphoblastoid and fibroblast cells (Le Tallec et al., 2011). In each cell type the expressed CFSs were located in regions devoid of replication initiation and at points of replication fork convergence. Furthermore, the expressed CFSs tend to be late replicating, indicating that the problems of fragility may occur at all CFSs but only persist to metaphase in those positions that have been unable to fully repair before cell division. This indicates that the DNA replication profile is likely to influence the expressed CFSs in a cell type dependent manner. However, the basis of the instability at CFSs which becomes exposed at late replicating regions by replication inhibition is unknown.

One hypothesis that could explain genome instability at late replicating, AT-rich regions is that replication introduces DNA supercoils which distort DNA structure, slow DNA polymerase and prevent the completion of DNA replication or assembly of metaphase chromosome structure prior to cell division. DNA polymerase introduces supercoils into the DNA that over-wind the helix ahead and under-wind both newly formed helices behind (Section 1.2.4.1.2). As two polymerases converge, over-wound DNA supercoils are introduced on an ever shorter length of DNA. The relief of these DNA supercoils may be further hampered at AT-rich sequences, which can absorb more DNA supercoils due to a higher flexibility in their structure (Burrow et al., 2010; Lukusa and Fryns, 2008; Le Tallec et al., 2011). In addition, the under-wound DNA upstream of each polymerase may stabilise alternative DNA structures or melt the DNA helix, which has been shown to stall RNA polymerase *in vitro* (Ma et al., 2013a). The build-up of over- and under-wound DNA supercoils associated with polymerases are usually relieved by the action of topoisomerase I and II, however at CFSs these enzymes are depleted (Section 3.2.7) which may further influence genome stability. Furthermore, the identification of stable over-wound DNA domains at FRA3B and FRA16D provides some indication that DNA supercoiling is involved at CFSs. To investigate whether DNA supercoil distribution changes at expressed CFSs it is necessary to map DNA supercoiling at CFSs with and without partial replication inhibition.

5.2 Results

5.2.1 RPE1 cells do not show fragility at FRA3B or FRA16D

The location and frequency of expressed CFSs varies between cell types, with FRA3B and FRA16D being the best characterised and most frequently expressed (Lukusa and Fryns, 2008; Le Tallec et al., 2011). However, the distribution of expressed CFSs in retinal pigmented epithelial cells (RPE1) is unknown. To identify appropriate conditions for the expression of CFSs by partial replication stress, cells were treated with three concentrations of aphidicolin and γH2AX foci quantified in interphase nuclei by immunofluorescence (Figure 5.2a). The number of cells with more than 5 foci did not dramatically increase with aphidicolin treatment (from 25% to 38%) (Figure 5.2b). This indicates that the majority of cells either do not form, or do not identify for repair, DNA breaks in interphase following aphidicolin treatment. In those cells with more than five foci, replication stress increases the number of γH2AX foci (Figure 5.2c). In the control cells, 35% of nuclei with γH2AX foci have more than 15 foci, whereas following 0.6 μM aphidicolin treatment this proportion increases to 92%. This indicates that in RPE1 cells aphidicolin increases γH2AX foci in cells that already show evidence of DNA damage, but hardly increases the proportion of cells with γH2AX foci contrary to observations in other cells (Schwartz et al., 2005). This may be due to differences in progression through the cell cycle, DNA damage repair mechanisms or chromatin structure between these cell lines. However, the presence of γH2AX foci on metaphase chromosomes indicates that DNA damage remains present at metaphase following 0.4 μM aphidicolin treatment (Figure 5.2a).

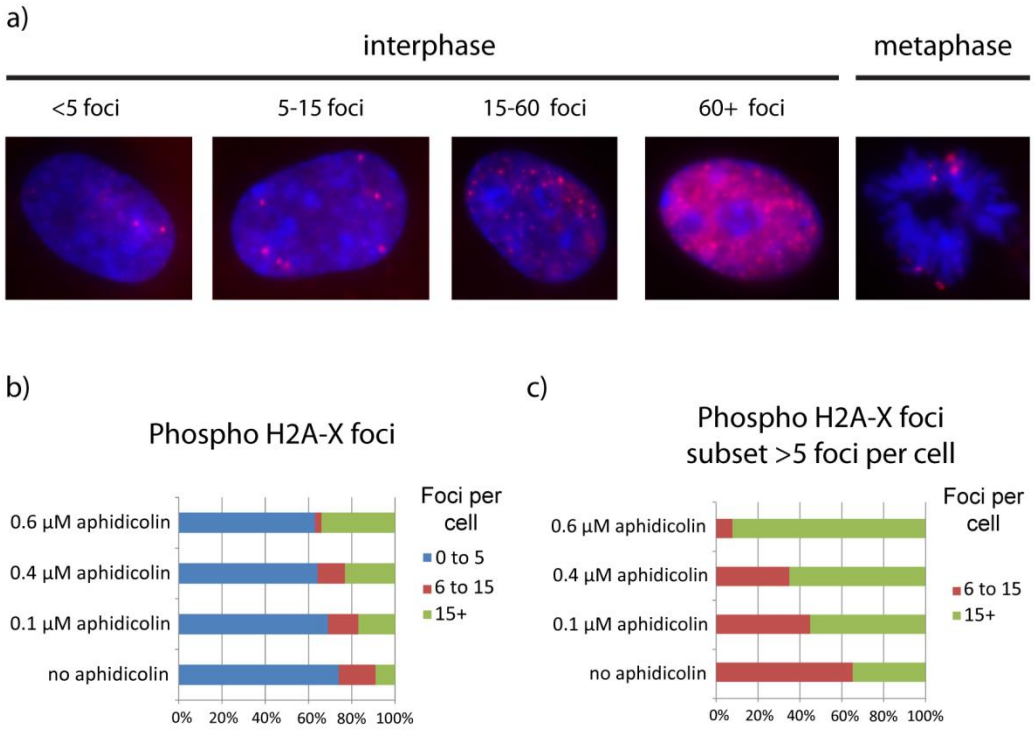


Figure 5.2 Replication inhibition increases DNA damage foci in RPE1 cells. a) Representative immunofluorescence images of γH2AX foci (red) in interphase and metaphase nuclei (DAPI – blue). b) Quantification of γH2AX foci in interphase cells following different aphidicolin treatments. Foci were quantified for 50 nuclei per condition. c) Quantification of foci in nuclei with damage following different aphidicolin treatments.

To identify if FRA3B and FRA16D are expressed CFSs in RPE1 cells, breaks were identified and mapped by karyotype analysis in giemsa treated metaphase chromosomes (Figure 5.3a). The two concentrations of aphidicolin used gave very different metaphase yields, with 0.6 µM aphidicolin producing almost no metaphases compared to 0.4 µM aphidicolin (data not shown). Using 0.4 µM aphidicolin, 12% of expressed CFSs were mapped to chromosome 3, although the majority of these breaks were not at FRA3B (Figure 5.3b). No expressed CFSs were mapped to chromosome 16. The most expressed CFSs were on chromosome 1, in particular at FRA1C (chromosome 1p31.2), indicating that RPE1 cells have a different expressed CFS profile to lymphoblastoid and fibroblast cells. Cells not treated with aphidicolin showed no expressed CFSs by karyotype analysis. Therefore, the topoisomerase, DNA supercoiling and other properties associated with FRA3B and FRA16D do not produce expressed CFSs in RPE1 cells.

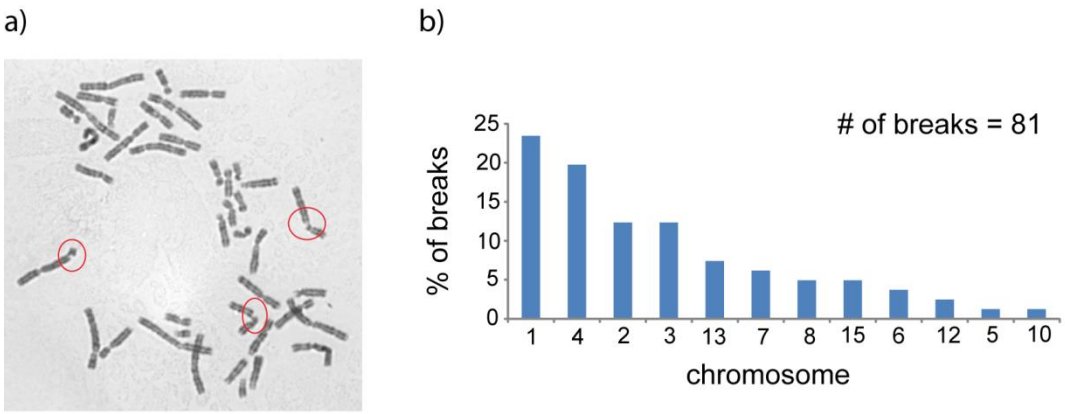


Figure 5.3 Karyotype analysis of RPE1 CFSs. a) Example metaphase for aphidicolin treat RPE1 cells showing 3 expressed CFSs (red circles). b) Proportion of CFSs on each chromosome for RPE1 cells. (Quantification performed by Christine Mordstein under my supervision)

5.2.2 Neo3 lymphoblastoid cells show fragility at FRA3B and FRA16D

To relate changes in DNA supercoiling at expressed CFSs to other established properties, an initial analysis should focus on the well-defined regions at FRA3B and FRA16D. As RPE1 cells do not show fragility at these CFSs several lymphoblastoid cell lines were assayed for good metaphase chromosome morphology and the appropriate expressed CFSs. Karyotype analysis of untreated metaphases (Figure 5.4) identified poor chromosome morphology for sweig and sato lymphoblastoid cell lines (e.g. sweig in Figure 5.4). These 'fuzzy' chromosomes were observed for several metaphase preparations, including a preparation alongside RPE1 chromosomes which gave good morphology. On the other hand, the neo3 lymphoblastoid cell line produced giemsa stained metaphase spreads with good chromosome morphology and were investigated further to identify which CFSs are expressed following partial replication inhibition.

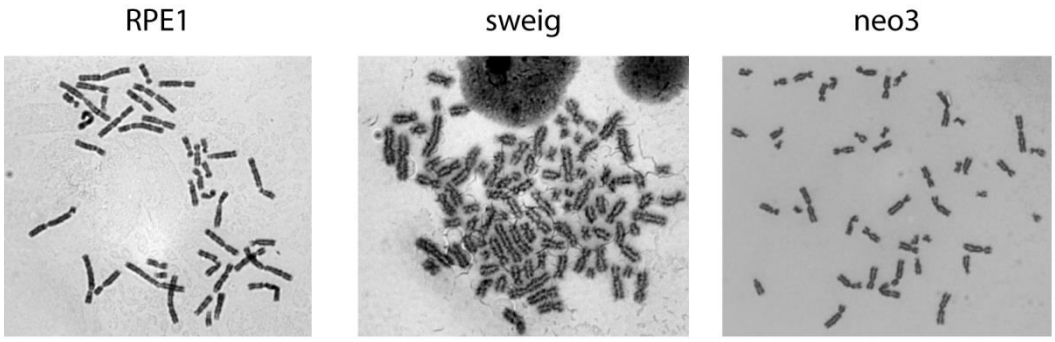


Figure 5.4 Metaphase morphology cell line comparison. Sweig cells have a poor chromosome morphology compared to RPE1 and neo3 cells.

To identify conditions of partial replication stress that induce DNA damage which may lead to CFS activation, neo3 cells were treated with three concentrations of aphidicolin. Following aphidicolin treatment there is a notable increase in the number of metaphases with more than 5 γH2AX foci, from 11% through to 58% (Figure 5.5a). This is a much higher proportion than observed for RPE1 cells (Section 5.2.1) and in line with other studies (Schwartz et al., 2005), indicating a higher level of aphidicolin induced DNA damage in neo3 cells. Analysis of giemsa stained chromosomes following 0.4 μM and 0.6 μM aphidicolin identifies a much higher proportion of chromosomes in metaphase than observed for RPE1 cells (data not shown), and allows the analysis of CFSs following 0.6 μM aphidicolin. The proportion of metaphases with breaks was roughly proportional to the proportion of interphase nuclei with greater than 5 γH2AX, 50% and 58% respectively. Mapping these breaks by karyotype analysis identifies frequent breaks on the chr3p (38%) and 16q (23%) arms, consistent with the positions of FRA3B and FRA16D (Figure 5.5b and 5.5c). Further confirmation of fragility at FRA3B was achieved by FISH analysis on metaphase chromosomes (Figure 5.5b). We were unable to distinguish between chromosomes 13, 14 and 15 through analysis of the neo3 karyotype, but it is likely that a single CFS occurs on one of these chromosomes and accounts for ~25% of breaks in aphidicolin treated neo3 metaphases. Therefore, consistent with reports in other lymphoblastoid cell lines, neo3 cells have expressed CFSs at FRA3B and FRA16D following partial replication inhibition.

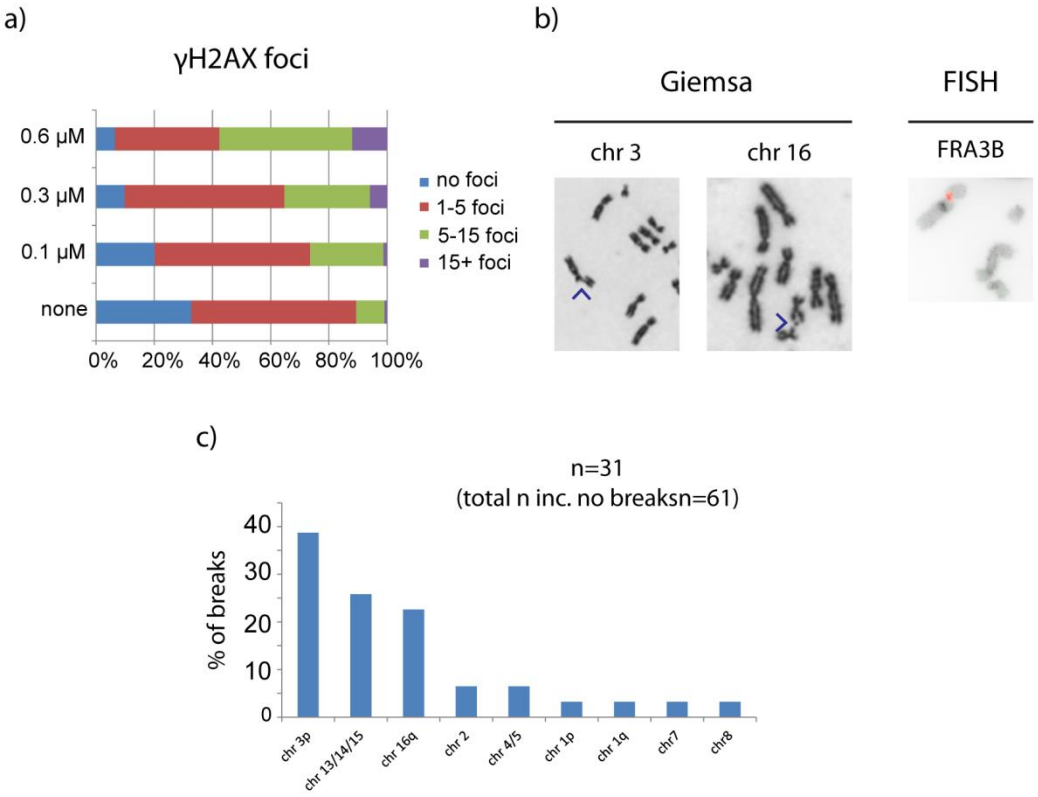


Figure 5.5 Mapping CFSs in neo3 cells. a) Quantification of γH2AX foci in interphase cells following different aphidicolin treatments (none - 113 nuclei, 0.1 μM – 238 nuclei, 0.3 μM – 153 nuclei, 0.6 μM – 92 nuclei). b) Example metaphase showing FRA3B and FRA16D CFSs by giemsa staining (arrows) and FRA3B CFS by FISH (red probe). c) Proportion of CFSs identified on chromosomes and/or chromosome arms. (FISH performed by Christine Mordstein)

5.2.3 bTMP pull-down identifies changes in DNA supercoiling at expressed CFSs

To identify whether changes in DNA supercoiling occur at expressed CFSs bTMP pull-down experiments were performed on neo3 cells, in the presence and absence of the DNA replication inhibitor aphidicolin, and hybridised to custom tiling microarrays (see section 2.7 for the design). As with previous experiments, a bTMP pull-down was performed on genomic DNA, as a control for the sequence preference of the drug. Furthermore, as a comparison with data for RPE1 cells a bTMP pull-down was performed on cells treated with the transcriptional inhibitor α-amanitin. Bioinformatic analysis of the microarray data was then performed to identify changes at the FRA3B and FRA16D CFSs.

To validate bTMP crosslink formation in non-adherent neo3 cells, a photo-crosslinking experiment was performed with two drug concentrations. At both concentrations bTMP is bound to the DNA following exposure to UV light, but not in the non-UV treated controls (Figure 5.6a). The level of photo-crosslinking in the high bTMP condition (1.4 mg/ml) is equivalent to one biotin every ~800 bp, which is similar to the previous analysis in RPE1 (one biotin every ~900 bp) (Section 4.2.1.4). Therefore, bTMP is cell-/nuclei- permeable and photo-crosslinks into DNA in neo3 cells, indicating that the bTMP pull-down technique described in Naughton et al. (2013) can be used in different cell types.

Further experimental validation confirms a reduction in DNA replication and transcription following inhibition with aphidicolin and α-amanitin respectively. Following aphidicolin treatment the cell cycle slowed, with a reduced number of cells in metaphase (data not shown). To confirm a reduction in transcription following α-amanitin treatment, cells were pulse labelled with tritiated uridine for 30 minutes and the relative incorporation of radioactivity into the RNA measured by scintillation counting. In α-amanitin treat cells there is a ~50% reduction in RNA production compared to control cells (Figure 5.6b). This incomplete loss of transcription may be accounted for by RNA polymerase I transcription, which is not inhibited at low doses of α-amanitin (Lindell et al., 1970). The inhibition of replication with aphidicolin also reduced RNA production, but only by ~20% (Figure

5.6b), possibly as a result of cell stress associated with replication inhibition. Therefore, this validation of bTMP photo-crosslinking and drug treatments indicates that these conditions in neo3 cells produce similar results to those observed in RPE1 cells (Naughton et al., 2013a) and should be appropriate for the identification of transcription and replication dependent DNA supercoil distributions.

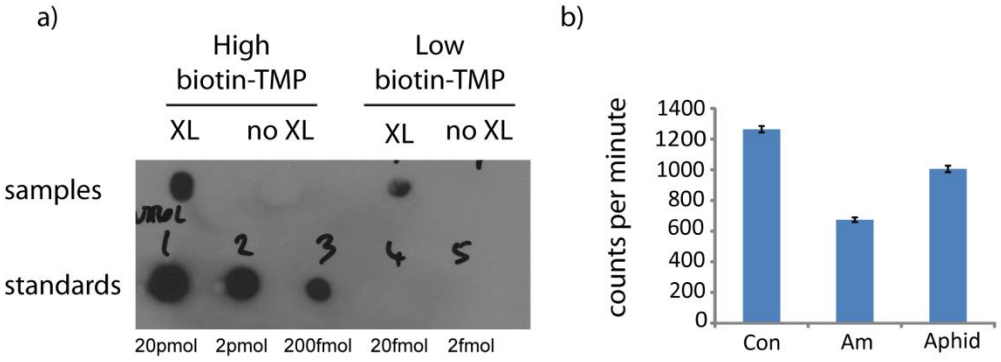


Figure 5.6 Validating bTMP pull-down experimental conditions. a) bTMP photo-crosslinking dotblot analysis. 500 ng DNA incubated with high (1400 µg/ml) and low (140 µg/ml) bTMP. The standards show the concentration of biotin. b) Tritiated uridine uptake as measure by scintillation counting identifies the relative RNA production over 30 minutes under different drug treatments. Counts per minute detects the rate of ionisation from the tritium label.

A bTMP pull-down was performed on neo3 cells and hybridised to microarrays tiling several loci including FRA3B and FRA16D. Microarray quality assessment and normalisation were performed (as in Section 3.2.3.1), using a VSN normalisation for consistency with Naughton et al. (2013a), and subsequent analysis was carried out on this data. An initial comparison between 'control – amanitin' bTMP distribution in neo3 and RPE1 cell lines identifies some similarity between DNA supercoiling in these cell types. For example, at the chromosome 11p15.5 locus (1 bp – 2800000 bp) of RPE1 and neo3 cells there is a general under-winding of the DNA with α-amanitin treatment over most of the locus, except at the most telomeric ~100 kb which becomes more over-wound (Figure 5.7). The IGBP1 locus on chromosome X (68369744 bp – 70369744 bp) shows some similarity in the 'control – amanitin' distribution between RPE1 and neo3 cells, with complementary regions of under-winding (Figure 5.7). However, the two distinct domains of over-winding observed in RPE1 cells are not present in the neo3 cells. This indicates that some DNA supercoil domains are cell type specific between RPE1 and neo3 cell lines. As a final example, the FRA3B locus on chromosome 3 (58600001 bp – 63700000 bp) again shows similarity between RPE1 and neo3 bTMP distribution, with a general depletion across the locus and similar regions of enrichment at several places, including the leftmost region of the plot (Figure 5.7).

The 'control – aphidicolin' distribution at these same loci identify no striking pattern of enrichment/depletion at the 11p15.5 or IGBP1 loci and look somewhat similar to the 'control – α-amanitin' distribution (Figure 5.7). This may be a reflection of the partial transcription inhibition observed following aphidicolin treatment (Figure 5.6). On the other hand, the FRA3B locus has a defined region of bTMP enrichment right of centre, indicative of a change in DNA supercoiling following DNA replication inhibition and the induction of CFSs.

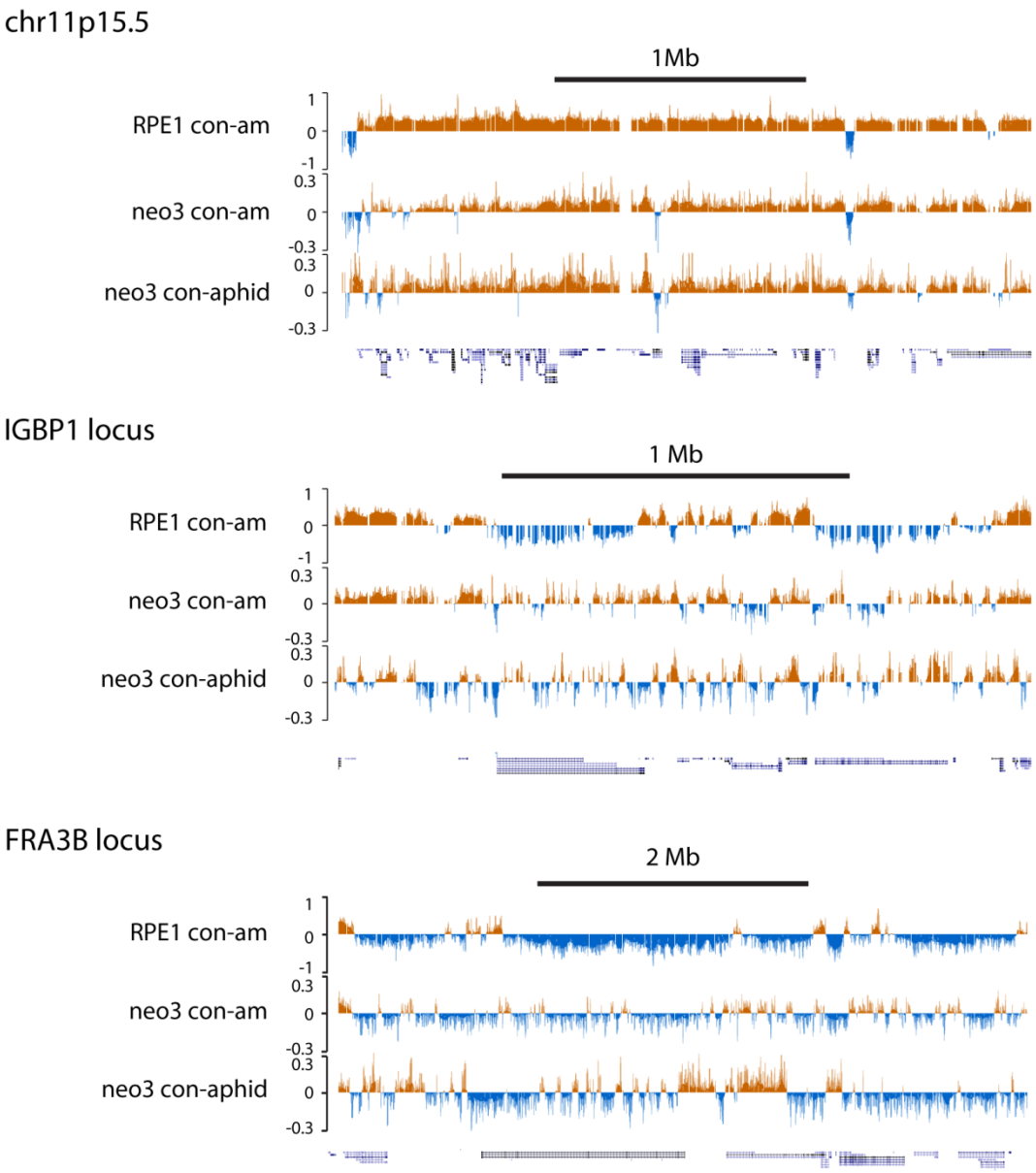


Figure 5.7 bTMP pull-down comparison between cell lines. Three regions from the custom tiling arrays covering chr11:1-2800000, chrX:68369744-70369744 and chr3:58600001-63700000 respectively. In each case the y-axis scale is either log2 (control / amanitin) or log2 (control / aphidicolin) bTMP enrichment over input. Genes are shown below the tracks.

To identify how DNA supercoiling changes following CFS expression in neo3 cells, the distribution of bTMP was analysed at FRA3B and FRA16D (Figure 5.8). The relationship between the FRA3B position defined in Becker et al. (2002), the topoisomerase depletion identified in RPE1 cells (Section 3.2.7) and an enrichment for bTMP in the 'control – aphidicolin' distribution (Figure 5.8) indicates a relationship between DNA supercoiling and the expression of CFSs following replication stress. Based on the distribution of these three factors, a core region was defined for FRA3B (Figure 5.8). The relationship between the FRA16D CFS as defined by Letessier et al. (2011) and an enrichment for 'control – aphidicolin' again suggests a relationship between CFS expression and DNA supercoiling. However, in this case the depletion of topoisomerase I in RPE1 cells is adjacent to the CFS position (Figure 5.8). Whether the depletion of topoisomerase I can effect CFS expression without their distributions overlapping, as shown for long genes (Le Tallec et al., 2013), or the distribution of topoisomerase I is different between RPE1 and neo3 cells is unknown. In the case of FRA16D, as the relationship with topoisomerase I is less clear the core region was defined based on the position of the cytologically mapped CFS. Together, this data indicates a relationship between DNA supercoiling and the expression of CFSs at FRA3B and FRA16D.

To further identify how DNA supercoiling changes at the FRA3B CFS following aphidicolin treatment the 'control – aphidicolin' distribution was compared to the 'control' and 'aphidicolin' distributions. There is a clear peak of bTMP in the 'control – aphidicolin' profile core region, particularly for FRA3B (Figure 5.8). The enrichment is significant and corresponds to a general shift to a more over-wound structure across the core region in samples treat with aphidicolin. Therefore, despite considerable variability over the FRA3B core, the under-wound regions become less-so and the over-wound regions become more over-wound generating the general shift to a more over-wound DNA structure. This indicates that at FRA3B, under conditions that induce visible CFSs at metaphase in ~20% of cells, the net change in DNA structure at the core region is over-wound compared to control.

To identify if a similar change in DNA supercoiling occurs at FRA16D, which induces visible CFSs at metaphase in ~10% of cells, a similar analysis of 'control –

aphidicolin', 'control' and 'aphidicolin' was performed. There is a significant difference between 'control – aphidicolin' bTMP between the core and flanking regions of the FRA16D CFS (Student's t test $p < 2.2x10^{-16}$). Following aphidicolin treatment the DNA within the CFS is generally more over-wound, similar to FRA3B. This indicates that partial replication inhibition leads to an over-winding of the DNA structure at these two CFSs and a relative under-winding of flanking sequences. Therefore, changes in DNA supercoiling occur at CFSs, supporting theoretical models in which DNA topology affects genome stability.

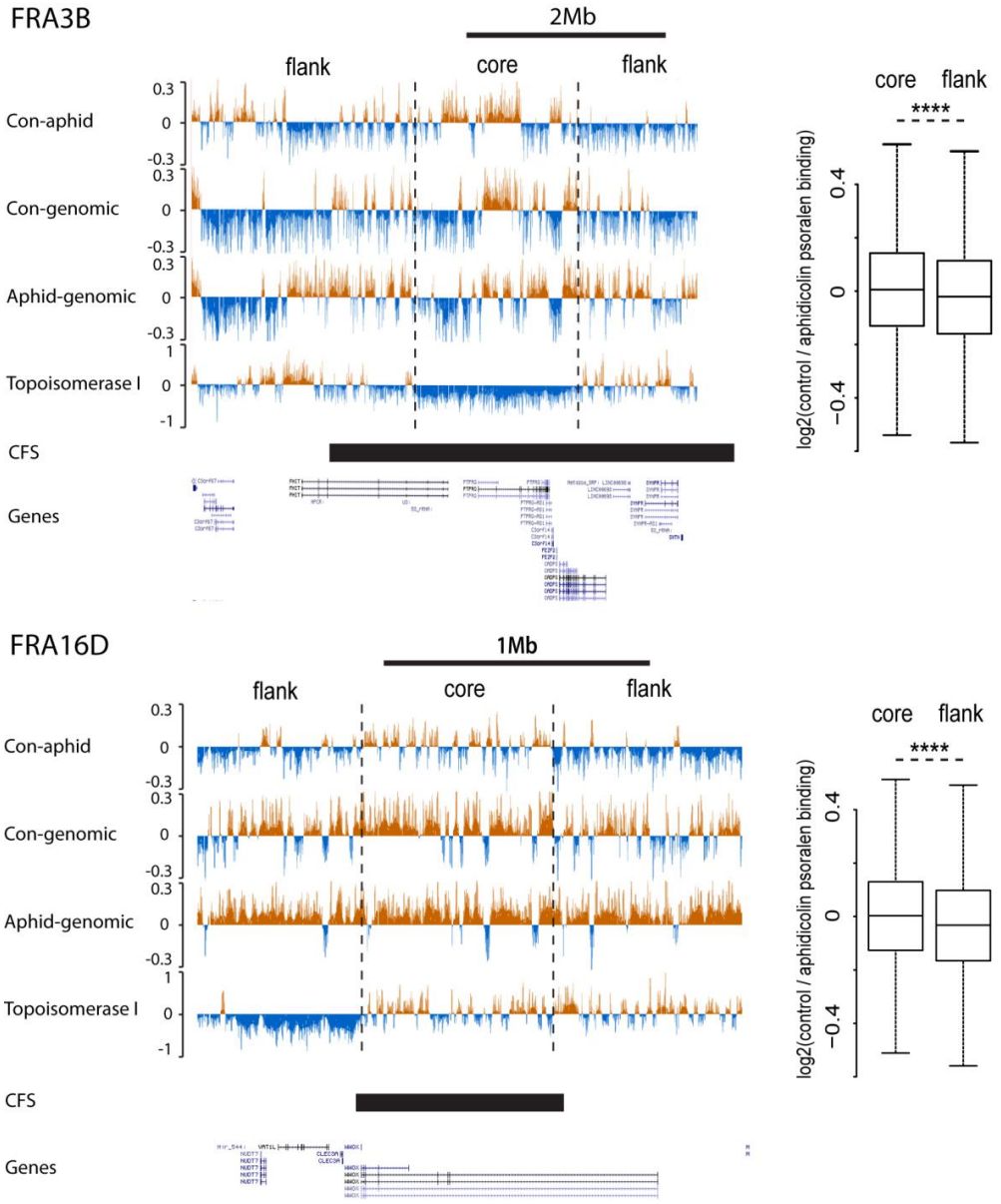


Figure 5.8 bTMP distribution at CFSs changes with partial replication inhibition. CFS regions are from Becker et al. (2002) (FRA3B) and Letessier et al. (2011) (FRA16D). Core refers to the region of the fragile site that is most commonly broken and flank to the surrounding sequence. Boxplots show the distribution of enrichment for 'control – aphidicolin' between the core and flanking regions. Students t test identifies a p value < $2.2x10^{-16}$ for FRA3B and FRA16D. Scale equals log2 (condition #1 / condition #2 bTMP enrichment) except topoisomerase which is log2 (topoisomerase I/input).

5.3 Discussion

The hypothesis that DNA supercoiling is important for genome stability at regions of flexible AT-rich DNA, including CFSs, is well supported by theoretical predictions (Section 5.1.1) but has not been determined experimentally. To identify *in vivo* if DNA supercoiling changes under conditions that activate CFSs, a bTMP pull-down was performed with and without aphidicolin treatment in lymphoblastoid cells. At the core of the expressed CFSs FRA3B and FRA16D a change in DNA supercoiling to a more over-wound DNA helix was observed following partial DNA replication inhibition, but not following transcription inhibition. In contrast, the DNA flanking these more over-wound regions becomes more under-wound following partial replication inhibition, indicating that the distribution of DNA supercoils varies considerably over the CFS loci. This data provides the first experimental evidence for a relationship between DNA supercoiling and the expression of CFSs *in vivo*.

Previous studies of DNA supercoiling at flexible regions of DNA have indicated that alternative DNA structures form following the introduction of under-wound DNA supercoils (Burrow et al., 2010; Gellibolian et al., 1997; Wojciechowska et al., 2006), and it has been proposed that these alternative DNA structures can slow the DNA polymerase and produce genome instability (Lukusa and Fryns, 2008). However, in the present study the DNA within defined CFSs is generally over-wound following partial replication inhibition, which will inhibit the DNA unwinding required for the formation of alternative DNA structures. This indicates a different mechanism for instability at CFSs, whereby over-wound DNA slows the replication machinery by inhibiting strand separation. The flanking DNA, which becomes more under-wound with aphidicolin treatment, may contribute to the strand slowing by a mechanism involving alternative DNA structures, but does not normally form the breakpoint. This supports a model whereby genome instability at CFSs results from a reduced ability to release over-wound DNA supercoils generated ahead of converging replication complexes, which is accentuated following replication inhibition. Sequence properties, topoisomerase depletions and converging replication forks at CFS loci can together reduce the efficiency of DNA

supercoil relaxation. The late replication of these regions, further slowed by aphidicolin, may cause chromosome breaks in metaphase due to the incomplete replication of DNA or the incomplete condensation of chromatin prior to cell division (Lukusa and Fryns, 2008).

Further analysis of DNA supercoiling at CFSs subsequent studies must apply the bTMP pull-down approach to analyse a much broader range of CFSs across several cell types. The relative change in DNA supercoiling across FRA3B and FRA16D is small, but significant. This may be due to the over-wound DNA structure occurring at different positions across the ~1 Mb region in a population of cells, consistent with the broad domain of genome instability. Therefore, an 'average' genome has a low level over-winding across a large domain composed of more significant over-wound structures at more localised regions of single genomes. It is not known whether over-winding occurs at all CFSs following partial replication inhibition or whether this is a property of expressed CFSs. To determine this, a bTMP pull-down must be performed on loci containing CFSs not expressed in a particular cells line, for example FRA16D in RPE1 cells. If changes in DNA supercoiling upon aphidicolin treatment occur only at expressed CFSs, then a meta-analysis of bTMP may be performed at expressed CFSs across a number of cell lines to identify whether a consistent structural change occurs. Either case will provide an important contribution in determining the role of DNA supercoiling at CFSs, with a presence only at expressed CFSs indicating a direct relationship with genome instability at these regions. Current work in the lab will focus on these experimental questions.

In addition to identifying for the first time changes in DNA supercoiling associated with CFSs, a number of other properties of CFSs and DNA supercoiling were identified. The distribution of CFSs in RPE1 cells was found to be distinct to those in lymphoblastoid, fibroblast and other epithelial cells, with FRA3C being the most expressed CFS (Le Tallec et al., 2011, 2013). This supports the observation of Le Tallec et al. (2011) that different cell types have different expressed CFSs and provides the first analysis of CFSs in RPE1 cells. The distribution of bTMP in the RPE1 and neo3 cell lines is also distinct at a number of the loci investigated. For example the depletions in bTMP observed in the RPE1 'control – amanitin'

distribution are not present in the equivalent neo3 distribution. Previous studies have indicated that DNA domains are relatively stable between cell lines and are conserved between species (Dixon et al., 2012). An attractive model for coordinated gene regulation is one in which domain boundaries remain constant, but the degree of DNA supercoiling within the boundaries varies with environmental conditions and/or cell type. A comparison between neo3 and RPE1 bTMP distribution provides tentative evidence for a cell-type dependent distribution of DNA supercoils at some loci. A more extensive analysis, such as that of the whole of chromosome 11 performed in (Naughton et al., 2013a), will provide important information regarding the role of DNA supercoiling in genome regulation.

In summary, this analysis has identified several crucial parameters of DNA supercoiling with relation to gene regulation and disease. By identifying similarities and differences in the distribution of bTMP between RPE1 and neo3 cells, the regulatory potential of this structural change is clarified. Furthermore, identifying changes in DNA supercoiling with the activation of CFSs at FRA3B and FRA16D supports experimentally the theoretical relationship between DNA structure and genome instability at these regions. Both of these observations form important pilot studies in the identification of the function of DNA supercoiling *in vivo*.

6. Conclusions

DNA supercoiling is important for regulating the expression of genes. For many years DNA supercoiling has been studied in plasmids and in the genomes of prokaryotes, but the distribution of unrestrained DNA supercoiling in the human genome was unknown. Recent work in our lab has identified large-scale DNA supercoiling domains in the human genome (Naughton et al., 2013a). Additionally, meta-analysis of several hundred gene promoters has identified an under-wound DNA supercoil distribution around the TSS (Kouzine et al., 2013; Naughton et al., 2013a). Both large-scale and promoter-scale distributions of DNA supercoiling are regulated by transcription and topoisomerase activity. However, the distribution of topoisomerases in the human genome remain poorly understood (Cowell et al., 2012; Khobta et al., 2006; Kouzine et al., 2013). In this thesis I have studied topoisomerases and DNA supercoiling to better characterise their relationship with gene expression and genome stability. Firstly, I have shown that topoisomerases form domain-scale and promoter-scale enrichments, in which topoisomerase I co-localises with RNA polymerase II and under-wound DNA whereas topoisomerase II co-localises with over-wound DNA. Secondly, I have characterised the distribution of DNA supercoiling at promoters genome-wide, identifying a more over-wound structure around the TSS of CpG island promoters which may be important for the regulation of these genes. Finally, I have identified changes in DNA supercoiling associated with CFS expression, an experimental observation which has not previously been made.

I have identified for the first time the distribution of topoisomerases in the human genome. The observation that topoisomerase I forms domains distinct from topoisomerase II suggests that two mechanisms of DNA supercoil release are required in different regions of the genome. Topoisomerase I is associated with RNA polymerase II and under-wound DNA, indicating a role in resolving the supercoils generated by transcription. This makes mechanistic sense, as the controlled rotation mechanism of topoisomerase I relaxes DNA supercoils in a manner dependent on the degree of supercoiling in the fibre (Koster et al., 2005).

Therefore, in regions with higher transcription activity topoisomerase I can rapidly release DNA supercoils without reaching saturation. Furthermore, topoisomerase I releases over-wound DNA fifty times faster than under-wound DNA (Frøhlich et al., 2007), which may contribute to the maintenance of the under-wound structure. Therefore, by virtue of its mechanism human topoisomerase I may both relax DNA supercoils and also contribute to the maintenance of an under-wound DNA structure, in a similar manner to prokaryotic topoisomerase IA (Kirkegaard and Wang, 1985). I have shown that the domain-scale distribution of topoisomerase II enzymes is distinct from that of topoisomerase I, but the distributions of topoisomerase IIα and IIβ are highly similar. Topoisomerase II domains are generally over-wound, which may facilitate the repression of genes (Ma et al., 2013b) or have other functional consequences in these regions. For example, the relationship between topoisomerase II enrichment at AT-rich, over-wound, gene poor DNA and the temporal distribution of topoisomerase IIα through the cell cycle (Woessner et al., 1991) may indicate that these domains are decatenation hotspots, taking advantage of the decatenation activity of topoisomerase II rather than its relaxase activity (Carpenter and Porter, 2004; McClendon et al., 2005; Woessner et al., 1991). Decatenation is critical in S through to G2-M and has been associated with both topoisomerase II and over-wound DNA, most notably in the decatenation of a yeast plasmid system (Baxter et al., 2011). Therefore, topoisomerase II domains may be decatenation hotspots, which occur in gene poor regions to limit the formation of potentially deleterious double strand breaks in coding regions of the genome.

In addition to a domain-scale distribution, I have identified that topoisomerases form a more local distribution at human gene promoters. A recent study suggested a general model of topoisomerase distribution, based on limited experimental data, in which topoisomerase I forms an invariant enrichment upstream of the TSS whereas topoisomerase II forms a peak of enrichment at the TSS in an expression dependent manner (Kouzine et al., 2013). My data suggests this model is incorrect, based on ChIP-chip experiments that determine the distribution of topoisomerase I and IIβ at high resolution around 2,509 TSSs. Topoisomerase I forms a peak at the TSS which is strongly enriched in expressed genes compared to non-expressed genes, whereas topoisomerase IIβ forms a peak that is independent of gene expression. Therefore, at

the promoter scale topoisomerase I is the critical topoisomerase for the maintenance of DNA supercoils. This agrees with my observations at a domain-scale, showing that topoisomerase I releases supercoils generated by transcription through a torsion dependent release mechanism at different scales in the genome. The function of topoisomerase II at gene promoters is less clear. A previous study in rat identified that topoisomerase II regulates gene expression in a minority of promoters (Sano et al., 2008) and in some cases this may occur at human promoters, but my analysis suggests that this is not a general mechanism for the regulation of DNA supercoiling at human gene promoters. Additionally, structural roles have been suggested for topoisomerase II which may account for some of the enrichment at promoters, including the association with chromatin remodelling complexes (Varga-Weisz et al., 1997).

Previous work has shown at a small sub-set of human gene promoters that a focal enrichment of under-wound DNA supercoiling occurs around the TSS (Kouzine et al., 2013; Naughton et al., 2013a). Using high density genome-wide promoter microarrays I have identified that a transcription dependent under-wound DNA structure is common at the promoter regions genome-wide. However, CpG island promoters show a strong depletion for under-wound DNA supercoiling in the ~1.5 kb surrounding the TSS which is not seen in non-CpG island promoters. As CpG islands make up the majority of human gene promoters an 'average' promoter genome-wide has this peak-trough-peak DNA supercoil distribution. Previous studies did not identify this structure, probably due to the use of data smoothing on the small number of genes analysed. My data suggests that the distribution of supercoiling at the TSS of CpG island promoters acts as a barrier to DNA supercoils and may impede strand separation. Therefore, as most CpG island promoters exist in an active chromatin environment, the relatively over-wound DNA structure may act as a general repressor of transcription initiation which can be released through DNA helicases (Singleton et al., 2007) or a transient increase in under-wound DNA supercoiling as a result of divergent transcription (Naughton et al., 2013b). Non-CpG island promoters do not have this structure at TSS, suggesting that a relatively over-wound TSS is not important for the regulation of these genes. As non-CpG island promoters are generally tissue-specific, it is likely that a general mechanism to

repress transcription initiation is unnecessary and these genes are instead regulated through the presence of tissue-specific transcription factors. I analysed promoters with respect to sequence and expression and identified that both are important for the structure of CpG island and non-CpG island promoters, but that neither one determines the DNA supercoil distribution alone. A comparison of expressed and non-expressed genes identifies that while CpG island promoters show a subtle change in DNA supercoil distribution, non-CpG island promoters show a strong rearrangement from an over-wound to an under-wound structure with expression. The DNA of non-expressed non-CpG island promoters has the most over-wound structure of any category of promoters analysed and may act as a general mechanism of transcription repression. The repressive role of relatively over-wound DNA supercoiling has not been discussed explicitly in previous studies and I present a model whereby over-wound DNA at the TSS of CpG island genes or across the promoter region of non-CpG island genes represses transcription initiation. In the case of CpG island genes, I propose that this repression is transiently released by helicase activity or the generation of additional supercoils by divergent transcription, whereas non-CpG island promoters give a more complete repression of expression. Further work is required to resolve the function of the relatively over-wound DNA I have observed at gene promoters *in vivo*. In addition, a comparison of generally expressed and generally repressed genes identified an expression independent under-wound DNA structure at genes which are expressed across a suite of tissue samples, but are not expressed in my cell line. This suggests that these promoters are being maintained in a paused supercoiling state, as observed in at a single locus in Ljungman and Hanawalt (1995). To identify how non-coding transcription may maintain an under-wound structure at expressed and non-expressed genes my current research is focusing on the relationship between DNA supercoiling, transcription and paused polymerase through an analysis of bTMP pull-down, CAGE and GRO-seq data.

The role of DNA supercoiling in genome stability is well established in the segregation of chromosomes during the cell cycle (Carpenter and Porter, 2004), but the predicted relationship between DNA supercoiling and chromosome instability at fragile sites has not been demonstrated experimentally. I have now shown that under

conditions of chromosome fragility at CFSs FRA3B and FRA16D the core regions become more over-wound compared to flanking DNA sequences. In addition, a strong depletion of both topoisomerase I and topoisomerase II is observed in the vicinity of FRA3B and FRA16D. This identifies for the first time a relationship between DNA supercoiling and CFS expression in human cells. In addition, I have shown that RPE1 cells have an expressed CFS profile distinct from lymphoblastoid, fibroblast, erythroid or other epithelial cells. The observation that FRA1C is the most expressed CFS in RPE1 cells, which is not expressed in five other epithelial cell types, highlights our poor understanding of the relationships between cell type and epigenetic inheritance of expressed CFS loci.

The results from my thesis suggest that DNA supercoiling plays an important role in gene expression and genome stability in human cells. I have characterised the inter-relationships between topoisomerases, RNA polymerase II and DNA supercoiling at domain- and promoter- scales and identified novel properties related to promoter structure and regulation. In addition, I have identified changes in DNA supercoiling at expressed CFS which support a role for DNA supercoiling in the stability of these regions. Future experiments will identify the relationship between transcription and DNA supercoil distribution, through an analysis of CAGE data, to complete the model of gene regulation through the maintenance of promoter DNA supercoil distribution. This study provides the basis for a complete understanding of the distribution, regulation and consequences of DNA supercoiling in the human genome.

7 References

Arents, G., Burlingame, R.W., Wang, B.C., Love, W.E., and Moudrianakis, E.N. (1991). The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. Proc. Natl. Acad. Sci. U.S.A. *88*, 10148–10152.

Bacolla, A., and Wells, R.D. (2009). Non-B DNA conformations as determinants of mutagenesis and human disease. Mol. Carcinog. *48*, 273–285.

Bacolla, A., Gellibolian, R., Shimizu, M., Amirhaeri, S., Kang, S., Ohshima, K., Larson, J.E., Harvey, S.C., Stollar, B.D., and Wells, R.D. (1997). Flexible DNA: genetically unstable CTG.CAG and CGG.CCG from human hereditary neuromuscular disease genes. J. Biol. Chem. *272*, 16783–16792.

Bak, A.L., Zeuthen, J., and Crick, F.H. (1977). Higher-order structure of human mitotic chromosomes. Proc. Natl. Acad. Sci. U.S.A. *74*, 1595–1599.

Bancaud, A., Conde e Silva, N., Barbi, M., Wagner, G., Allemand, J.-F., Mozziconacci, J., Lavelle, C., Croquette, V., Victor, J.-M., Prunell, A., et al. (2006). Structural plasticity of single chromatin fibers revealed by torsional manipulation. Nat. Struct. Mol. Biol. *13*, 444–450.

Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res. *21*, 381–395.

Bates, A.D., and Maxwell, A. (2005). DNA topology (Oxford; New York: Oxford University Press).

Baxter, J., and Aragón, L. (2010). Physical linkages between sister chromatids and their removal during yeast chromosome segregation. Cold Spring Harb. Symp. Quant. Biol. *75*, 389–394.

Baxter, J., Sen, N., Martínez, V.L., De Carandini, M.E.M., Schvartzman, J.B., Diffley, J.F.X., and Aragón, L. (2011). Positive supercoiling of mitotic DNA drives decatenation by topoisomerase II in eukaryotes. Science *331*, 1328–1332.

Becker, N.A., Thorland, E.C., Denison, S.R., Phillips, L.A., and Smith, D.I. (2002). Evidence that instability within the FRA3B region extends four megabases. Oncogene *21*, 8713–8722.

Behe, M., and Felsenfeld, G. (1981). Effects of methylation on a synthetic polynucleotide: the B--Z transition in poly(dG-m5dC).poly(dG-m5dC). Proc. Natl. Acad. Sci. U.S.A. *78*, 1619–1623.

Belmont, A.S., and Bruce, K. (1994). Visualization of G1 chromosomes: a folded, twisted, supercoiled chromonema model of interphase chromatid structure. J. Cell Biol. *127*, 287–302.

Benyajati, C., and Worcel, A. (1976). Isolation, characterization, and structure of the folded interphase genome of Drosophila melanogaster. Cell *9*, 393–407.

Bergerat, A., Gadelle, D., and Forterre, P. (1994). Purification of a DNA topoisomerase II from the hyperthermophilic archaeon Sulfolobus shibatae. A thermostable enzyme with both bacterial and eucaryal features. J. Biol. Chem. *269*, 27663–27669.

Bermúdez, I., García-Martínez, J., Pérez-Ortín, J.E., and Roca, J. (2010). A method for genome-wide analysis of DNA helical tension by means of psoralen-DNA photobinding. Nucleic Acids Res. *38*, e182.

Bickmore, W.A., and Van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*, 1270–1284.

Biffi, G., Tannahill, D., McCafferty, J., and Balasubramanian, S. (2013). Quantitative visualization of DNA G-quadruplex structures in human cells. Nat Chem *5*, 182–186.

Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209–213.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell *40*, 91–99.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. (1997). The complete genome sequence of Escherichia coli K-12. Science *277*, 1453–1462.

Boyle, S., Rodesch, M.J., Halvensleben, H.A., Jeddeloh, J.A., and Bickmore, W.A. (2011). Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosome Res. *19*, 901–909.

Branco, M.R., and Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol. *4*, e138.

Branello, L., Levens, D., Gupta, A., and Kouzine, F. (2012). The importance of being supercoiled: how DNA mechanics regulate dynamic processes. Biochimica Et Biophysica Acta *1819*.

Branham, W.S., Melvin, C.D., Han, T., Desai, V.G., Moland, C.L., Scully, A.T., and Fuscoe, J.C. (2007). Elimination of laboratory ozone leads to a dramatic improvement in the reproducibility of microarray gene expression measurements. BMC Biotechnol. *7*, 8.

Brázda, V., Laister, R.C., Jagelská, E.B., and Arrowsmith, C. (2011). Cruciform structures are a common DNA feature important for regulating biological processes. BMC Mol. Biol. *12*, 33.

Burrow, A.A., Marullo, A., Holder, L.R., and Wang, Y.-H. (2010). Secondary structure formation and DNA instability at fragile site FRA16B. Nucleic Acids Res. *38*, 2865–2877.

Cairns, B.R. (2009). The logic of chromatin architecture and remodelling at promoters. Nature *461*, 193–198.

Calladine, C.R. (2004). Understanding DNA the molecule & how it works. (Amsterdam; London: Elsevier Academic Press).

Carpenter, A.J., and Porter, A.C.G. (2004). Construction, characterization, and complementation of a conditional-lethal DNA topoisomerase IIalpha mutant human cell line. Mol. Biol. Cell *15*, 5700–5711.

Cech, T., and Pardue, M.L. (1977). Cross-linking of DNA with trimethylpsoralen is a probe for chromatin structure. Cell *11*, 631–640.

Cer, R.Z., Bruce, K.H., Mudunuri, U.S., Yi, M., Volfovsky, N., Luke, B.T., Bacolla, A., Collins, J.R., and Stephens, R.M. (2011). Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. Nucleic Acids Res. *39*, D383–391.

Champoux, J.J. (2001). DNA topoisomerases: structure, function, and mechanism. Annu. Rev. Biochem. *70*, 369–413.

Champoux, J.J., and Dulbecco, R. (1972). An activity from mammalian cells that untwists superhelical DNA--a possible swivel for DNA replication (polyoma-ethidium bromide-mouse-embryo cells-dye binding assay). Proc. Natl. Acad. Sci. U.S.A. *69*, 143–146.

Chan, K.-L., North, P.S., and Hickson, I.D. (2007). BLM is required for faithful chromosome segregation and its localization defines a class of ultrafine anaphase bridges. EMBO J. *26*, 3397–3409.

Chang, C.-W., Cheng, W.-C., Chen, C.-R., Shu, W.-Y., Tsai, M.-L., Huang, C.-L., and Hsu, I.C. (2011). Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. PLoS ONE *6*, e22859.

Christensen, M.O., Krokowski, R.M., Barthelmes, H.U., Hock, R., Boege, F., and Mielke, C. (2004). Distinct effects of topoisomerase I and RNA polymerase I inhibitors suggest a dual mechanism of nucleolar/nucleoplasmic partitioning of topoisomerase I. J. Biol. Chem. *279*, 21873–21882.

Cimino, G.D., Gamper, H.B., Isaacs, S.T., and Hearst, J.E. (1985). Psoralens as photoactive probes of nucleic acid structure and function: organic chemistry, photochemistry, and biochemistry. Annu. Rev. Biochem. *54*, 1151–1193.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

Corona, D.F.V., and Tamkun, J.W. (2004). Multiple roles for ISWI in transcription, chromosome organization and DNA replication. Biochim. Biophys. Acta *1677*, 113–119.

Costanzi, C., and Pehrson, J.R. (1998). Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals. Nature *393*, 599–601.

Cowell, I.G., Sondka, Z., Smith, K., Lee, K.C., Manville, C.M., Sidorczuk-Lesthuruge, M., Rance, H.A., Padget, K., Jackson, G.H., Adachi, N., et al. (2012). Model for MLL translocations in therapy-related leukemia involving topoisomerase IIβ-mediated DNA strand breaks and gene proximity. Proc. Natl. Acad. Sci. U.S.A. *109*, 8989–8994.

Crisona, N.J., Strick, T.R., Bensimon, D., Croquette, V., and Cozzarelli, N.R. (2000). Preferential relaxation of positively supercoiled DNA by E. coli topoisomerase IV in single-molecule and ensemble measurements. Genes Dev. *14*, 2881–2892.

Davey, C.A., Sargent, D.F., Luger, K., Maeder, A.W., and Richmond, T.J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. J. Mol. Biol. *319*, 1097–1113.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Desai, S.D., Liu, L.F., Vazquez-Abad, D., and D'Arpa, P. (1997). Ubiquitin-dependent destruction of topoisomerase I is stimulated by the antitumor drug camptothecin. J. Biol. Chem. *272*, 24159–24164.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dodge, J.E., Kang, Y.-K., Beppu, H., Lei, H., and Li, E. (2004). Histone H3-K9 methyltransferase ESET is essential for early development. Mol. Cell. Biol. *24*, 2478–2486.

Drew, H.R., and Travers, A.A. (1985). DNA bending and its relation to nucleosome positioning. J. Mol. Biol. *186*, 773–790.

Dunaway, M., and Ostrander, E.A. (1993). Local domains of supercoiling activate a eukaryotic promoter in vivo. Nature *361*, 746–748.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Durand-Dubief, M., Persson, J., Norman, U., Hartsuiker, E., and Ekwall, K. (2010). Topoisomerase I regulates open chromatin and controls gene expression in vivo. EMBO J. *29*, 2126–2134.

Earnshaw, W.C., and Heck, M.M. (1985). Localization of topoisomerase II in mitotic chromosomes. J. Cell Biol. *100*, 1716–1725.

Ebisuya, M., Yamamoto, T., Nakajima, M., and Nishida, E. (2008). Ripples from neighbouring transcription. Nat. Cell Biol. *10*, 1106–1113.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res. *10*, 2709–2721.

Eisenberg, E., and Levanon, E.Y. (2003). Human housekeeping genes are compact. Trends Genet. *19*, 362–365.

Engle, E.C., Manes, S.H., and Drlica, K. (1982). Differential effects of antibiotics inhibiting gyrase. J. Bacteriol. *149*, 92–98.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Esposito, F., Brankamp, R.G., and Sinden, R.R. (1988). DNA sequence specificity of 4,5',8-trimethylpsoralen cross-linking. Effect of neighboring bases on cross-linking the 5'-TA dinucleotide. J. Biol. Chem. *263*, 11466–11472.

Filion, G.J., Van Bemmel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., De Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J., et al. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell *143*, 212–224.

FitzGerald, P.C., and Simpson, R.T. (1985). Effects of sequence alterations in a DNA segment containing the 5 S RNA gene from Lytechinus variegatus on positioning of a nucleosome core particle in vitro. J. Biol. Chem. *260*, 15318–15324.

Forterre, P., and Gadelle, D. (2009). Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms. Nucleic Acids Res. *37*, 679–692.

French, S.L., Sikes, M.L., Hontz, R.D., Osheim, Y.N., Lambert, T.E., El Hage, A., Smith, M.M., Tollervey, D., Smith, J.S., and Beyer, A.L. (2011). Distinguishing the roles of Topoisomerases I and II in relief of transcription-induced torsional stress in yeast rRNA genes. Mol. Cell. Biol. *31*, 482–494.

Frøhlich, R.F., Veigaard, C., Andersen, F.F., McClendon, A.K., Gentry, A.C., Andersen, A.H., Osheroff, N., Stevnsner, T., and Knudsen, B.R. (2007). Tryptophane-205 of human topoisomerase I is essential for camptothecin inhibition of negative but not positive supercoil removal. Nucleic Acids Res. *35*, 6170–6180.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

Fussner, E., Ching, R.W., and Bazett-Jones, D.P. (2011). Living without 30nm chromatin fibers. Trends Biochem. Sci. *36*, 1–6.

Gabrielli, A., Avvedimento, E.V., and Krieg, T. (2009). Scleroderma. N. Engl. J. Med. *360*, 1989–2003.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.

Gellert, M., Mizuuchi, K., O'Dea, M.H., and Nash, H.A. (1976). DNA gyrase: an enzyme that introduces superhelical turns into DNA. Proc. Natl. Acad. Sci. U.S.A. *73*, 3872–3876.

Gellibolian, R., Bacolla, A., and Wells, R.D. (1997). Triplet repeat instability and DNA topology: an expansion model based on statistical mechanics. J. Biol. Chem. *272*, 16793–16797.

Giaever, G.N., and Wang, J.C. (1988). Supercoiling of intracellular DNA can occur in eukaryotic cells. Cell *55*, 849–856.

Gilbert, N., and Allan, J. (2001). Distinctive higher-order chromatin structure at mammalian centromeres. Proc. Natl. Acad. Sci. U.S.A. *98*, 11949–11954.

Gilbert, N., and Allan, J. (2013). Supercoiling in DNA and chromatin. Curr. Opin. Genet. Dev. *25C*, 15–21.

Gilmour, D.S., Pflugfelder, G., Wang, J.C., and Lis, J.T. (1986). Topoisomerase I interacts with transcribed regions in Drosophila cells. Cell *44*, 401–407.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., De Klein, A., Wessels, L., De Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.

Guldner, H.H., Szostecki, C., Vosberg, H.P., Lakomek, H.J., Penner, E., and Bautz, F.A. (1986). Scl 70 autoantibodies from scleroderma patients recognize a 95 kDa protein identified as DNA topoisomerase I. Chromosoma *94*, 132–138.

Hamdouch, K., Rodríguez, C., Pérez-Venegas, J., Rodríguez, I., Astola, A., Ortiz, M., Yen, T.J., Bennani, M., and Valdivia, M.M. (2011). Anti-CENPI autoantibodies in scleroderma patients with features of autoimmune liver diseases. Clin. Chim. Acta *412*, 2267–2271.

El Hassan, M.A., and Calladine, C.R. (1996). Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. J. Mol. Biol. *259*, 95–103.

Hayes, J.J., Clark, D.J., and Wolffe, A.P. (1991). Histone contributions to the structure of DNA in the nucleosome. Proc. Natl. Acad. Sci. U.S.A. *88*, 6829–6833.

Hays, F.A., Teegarden, A., Jones, Z.J.R., Harms, M., Raup, D., Watson, J., Cavaliere, E., and Ho, P.S. (2005). How sequence defines structure: a crystallographic map of DNA structure and conformation. Proc. Natl. Acad. Sci. U.S.A. *102*, 7157–7162.

He, Y., Fang, J., Taatjes, D.J., and Nogales, E. (2013). Structural visualization of key steps in human transcription initiation. Nature *495*, 481–486.

Helmrich, A., Stout-Weider, K., Hermann, K., Schrock, E., and Heiden, T. (2006). Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes. Genome Res. *16*, 1222–1230.

Henikoff, S., Henikoff, J.G., Sakai, A., Loeb, G.B., and Ahmad, K. (2009). Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome Res. *19*, 460–469.

Herbert, A., Lowenhaupt, K., Spitzner, J., and Rich, A. (1995). Chicken double-stranded RNA adenosine deaminase has apparent specificity for Z-DNA. Proc. Natl. Acad. Sci. U.S.A. *92*, 7550–7554.

Herbert, A., Alfken, J., Kim, Y.G., Mian, I.S., Nishikura, K., and Rich, A. (1997). A Z-DNA binding domain present in the human editing enzyme, double-stranded RNA adenosine deaminase. Proc. Natl. Acad. Sci. U.S.A. *94*, 8421–8426.

Hiasa, H., DiGate, R.J., and Marians, K.J. (1994). Decatenating activity of Escherichia coli DNA gyrase and topoisomerases I and III during oriC and pBR322 DNA replication in vitro. J. Biol. Chem. *269*, 2093–2099.

Hirose, S., and Suzuki, Y. (1988). In vitro transcription of eukaryotic genes is affected differently by the degree of DNA supercoiling. Proc. Natl. Acad. Sci. U.S.A. *85*, 718–722.

Hogan, M.E., Rooney, T.F., and Austin, R.H. (1987). Evidence for kinks in DNA folding in the nucleosome. Nature *328*, 554–557.

Van Holde, K.E. (1989). Chromatin (New York: Springer-Verlag).

Hon, G., Wang, W., and Ren, B. (2009). Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput. Biol. *5*, e1000566.

Howman, E.V., Fowler, K.J., Newson, A.J., Redward, S., MacDonald, A.C., Kalitsis, P., and Choo, K.H. (2000). Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. Proc. Natl. Acad. Sci. U.S.A. *97*, 1148–1153.

Huang, Y., Fang, J., Bedford, M.T., Zhang, Y., and Xu, R.-M. (2006). Recognition of histone H3 lysine-4 methylation by the double tudor domain of JMJD2A. Science *312*, 748–751.

Hughes, A.L., Jin, Y., Rando, O.J., and Struhl, K. (2012). A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern. Mol. Cell *48*, 5–15.

Irizarry, R.A., Wu, H., and Feinberg, A.P. (2009). A species-generalized probabilistic model-based definition of CpG islands. Mamm. Genome *20*, 674–680.

Isik, S., Sano, K., Tsutsui, K., Seki, M., Enomoto, T., Saitoh, H., and Tsutsui, K. (2003). The SUMO pathway is required for selective degradation of DNA topoisomerase IIbeta induced by a catalytic inhibitor ICRF-193(1). FEBS Lett. *546*, 374–378.

Jabs, E.W., Tuck-Muller, C.M., Anhalt, G.J., Earnshaw, W., Wise, R.A., and Wigley, F. (1993). Cytogenetic survey in systemic sclerosis: correlation of aneuploidy with the presence of anticentromere antibodies. Cytogenet. Cell Genet. *63*, 169–175.

Jeong, K.S., Ahn, J., and Khodursky, A.B. (2004). Spatial patterns of transcriptional activity in the chromosome of Escherichia coli. Genome Biol. *5*, R86.

Johnstone, A. (1996). Immunochemistry in practice (Cambridge, Mass: Blackwell Science).

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

Joshi, R.S., Piña, B., and Roca, J. (2010). Positional dependence of transcriptional inhibition by DNA torsional stress in yeast chromosomes. EMBO J. *29*, 740–748.

Ju, B.-G., Lunyak, V.V., Perissi, V., Garcia-Bassets, I., Rose, D.W., Glass, C.K., and Rosenfeld, M.G. (2006). A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription. Science *312*, 1798–1802.

Jupe, E.R., Sinden, R.R., and Cartwright, I.L. (1993). Stably maintained microdomain of localized unrestrained supercoiling at a Drosophila heat shock gene locus. EMBO J. *12*, 1067–1075.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90–98.

Kampmann, M., and Stock, D. (2004). Reverse gyrase has heat-protective DNA chaperone activity independent of supercoiling. Nucleic Acids Res. *32*, 3537–3545.

Kanne, D., Straub, K., Rapoport, H., and Hearst, J.E. (1982). Psoralen-deoxyribonucleic acid photoreaction. Characterization of the monoaddition products from 8-methoxypsoralen and 4,5’8-trimethylpsoralen. Biochemistry *21*, 861–871.

Käs, E., and Laemmli, U.K. (1992). In vivo topoisomerase II cleavage of the Drosophila histone and satellite III repeats: DNA sequence and structural characteristics. EMBO J. *11*, 705–716.

Kavenoff, R., and Ryder, O.A. (1976). Electron microscopy of membrane-associated folded chromosomes of Escherichia coli. Chromosoma *55*, 13–25.

Keshet, I., Yisraeli, J., and Cedar, H. (1985). Effect of regional DNA methylation on gene expression. Proc. Natl. Acad. Sci. U.S.A. *82*, 2560–2564.

Khobta, A., Ferri, F., Lotito, L., Montecucco, A., Rossi, R., and Capranico, G. (2006). Early effects of topoisomerase I inhibition on RNA polymerase II along transcribed genes in human cells. J. Mol. Biol. *357*, 127–138.

Kim, R.A., and Wang, J.C. (1989). Function of DNA topoisomerases as replication swivels in Saccharomyces cerevisiae. J. Mol. Biol. *208*, 257–267.

Kim, J.L., Nikolov, D.B., and Burley, S.K. (1993). Co-crystal structure of TBP recognizing the minor groove of a TATA element. Nature *365*, 520–527.

Kirkegaard, K., and Wang, J.C. (1985). Bacterial DNA topoisomerase I can relax positively supercoiled DNA containing a single-stranded loop. J. Mol. Biol. *185*, 625–637.

Koster, D.A., Croquette, V., Dekker, C., Shuman, S., and Dekker, N.H. (2005). Friction and torque govern the relaxation of DNA supercoils by eukaryotic topoisomerase IB. Nature *434*, 671–674.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693–705.

Kouzine, F., Liu, J., Sanford, S., Chung, H.-J., and Levens, D. (2004). The dynamic response of upstream DNA to transcription-generated torsional stress. Nat. Struct. Mol. Biol. *11*, 1092–1100.

Kouzine, F., Sanford, S., Elisha-Feil, Z., and Levens, D. (2008). The functional response of upstream DNA to dynamic supercoiling in vivo. Nat. Struct. Mol. Biol. *15*, 146–154.

Kouzine, F., Gupta, A., Baranello, L., Wojtowicz, D., Ben-Aissa, K., Liu, J., Przytycka, T.M., and Levens, D. (2013). Transcription-dependent dynamic supercoiling is a short-range genomic force. Nat. Struct. Mol. Biol. *20*, 396–403.

Kozyavkin, S.A., Pushkin, A.V., Eiserling, F.A., Stetter, K.O., Lake, J.A., and Slesarev, A.I. (1995). DNA enzymology above 100 degrees C. Topoisomerase V unlinks circular DNA at 80-122 degrees C. J. Biol. Chem. *270*, 13593–13595.

Kretzschmar, M., Meisterernst, M., and Roeder, R.G. (1993). Identification of human DNA topoisomerase I as a cofactor for activator-dependent transcription by RNA polymerase II. Proc. Natl. Acad. Sci. U.S.A. *90*, 11508–11512.

Kruithof, M., Chien, F.-T., Routh, A., Logie, C., Rhodes, D., and Van Noort, J. (2009). Single-molecule force spectroscopy reveals a highly compliant helical folding for the 30-nm chromatin fiber. Nat. Struct. Mol. Biol. *16*, 534–540.

Ku, M., Jaffe, J.D., Koche, R.P., Rheinbay, E., Endoh, M., Koseki, H., Carr, S.A., and Bernstein, B.E. (2012). H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. Genome Biol. *13*, R85.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950–953.

Lee, M.P., Brown, S.D., Chen, A., and Hsieh, T.S. (1993). DNA topoisomerase I is essential in Drosophila melanogaster. Proc. Natl. Acad. Sci. U.S.A. *90*, 6656–6660.

Lee, T.I., Johnstone, S.E., and Young, R.A. (2006). Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc *1*, 729–748.

Leppard, J.B., and Champoux, J.J. (2005). Human DNA topoisomerase I: relaxation, roles, and damage control. Chromosoma *114*, 75–85.

Letessier, A., Millot, G.A., Koundrioukoff, S., Lachagès, A.-M., Vogt, N., Hansen, R.S., Malfoy, B., Brison, O., and Debatisse, M. (2011). Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. Nature *470*, 120–123.

Li, Z., Mondragón, A., Hiasa, H., Marians, K.J., and DiGate, R.J. (2000). Identification of a unique domain essential for Escherichia coli DNA topoisomerase III-catalysed decatenation of replication intermediates. Mol. Microbiol. *35*, 888–895.

Li, Z., Hiasa, H., and DiGate, R. (2005). Bacillus cereus DNA topoisomerase I and IIIalpha: purification, characterization and complementation of Escherichia coli TopoIII activity. Nucleic Acids Res. *33*, 5415–5425.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lindell, T.J., Weinberg, F., Morris, P.W., Roeder, R.G., and Rutter, W.J. (1970). Specific inhibition of nuclear RNA polymerase II by alpha-amanitin. Science *170*, 447–449.

Lipps, H.J., and Rhodes, D. (2009). G-quadruplex structures: in vivo evidence and function. Trends Cell Biol. *19*, 414–422.

Liu, L.F., and Wang, J.C. (1979). Interaction between DNA and Escherichia coli DNA topoisomerase I. Formation of complexes between the protein and superhelical and nonsuperhelical duplex DNAs. J. Biol. Chem. *254*, 11082–11088.

Liu, L.F., and Wang, J.C. (1987). Supercoiling of the DNA template during transcription. Proc. Natl. Acad. Sci. U.S.A. *84*, 7024–7027.

Liu, L.F., Liu, C.C., and Alberts, B.M. (1980). Type II DNA topoisomerases: enzymes that can unknot a topologically knotted DNA molecule via a reversible double-strand break. Cell *19*, 697–707.

Ljungman, M., and Hanawalt, P.C. (1992). Localized torsional tension in the DNA of human cells. Proc. Natl. Acad. Sci. U.S.A. *89*, 6055–6059.

Ljungman, M., and Hanawalt, P.C. (1995). Presence of negative torsional tension in the promoter region of the transcriptionally poised dihydrofolate reductase gene in vivo. Nucleic Acids Res. *23*, 1782–1789.

Lu, K., Ye, W., Zhou, L., Collins, L.B., Chen, X., Gold, A., Ball, L.M., and Swenberg, J.A. (2010). Structural characterization of formaldehyde-induced cross-links between amino acids and deoxynucleosides and their oligomers. J. Am. Chem. Soc. *132*, 3388–3399.

Lu, X.J., Shakked, Z., and Olson, W.K. (2000). A-form conformational motifs in ligand-bound DNA structures. J. Mol. Biol. *300*, 819–840.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251–260.

Lukusa, T., and Fryns, J.P. (2008). Human chromosome fragility. Biochim. Biophys. Acta *1779*, 3–16.

Lyu, Y.L., Lin, C.-P., Azarova, A.M., Cai, L., Wang, J.C., and Liu, L.F. (2006). Role of topoisomerase IIbeta in the expression of developmentally regulated genes. Mol. Cell. Biol. *26*, 7929–7941.

Ma, J., Bai, L., and Wang, M.D. (2013a). Transcription under torsion. Science *340*, 1580–1583.

Ma, J., Bai, L., and Wang, M.D. (2013b). Transcription under torsion. Science *340*, 1580–1583.

Madden, K.R., Stewart, L., and Champoux, J.J. (1995). Preferential binding of human topoisomerase I to superhelical DNA. EMBO J. *14*, 5399–5409.

Maeshima, K., Hihara, S., and Eltsov, M. (2010). Chromatin structure: does the 30-nm fibre exist in vivo? Curr. Opin. Cell Biol. *22*, 291–297.

Martens, J.A., and Winston, F. (2003). Recent advances in understanding chromatin remodeling by Swi/Snf complexes. Curr. Opin. Genet. Dev. *13*, 136–142.

Martínez-Robles, M.L., Witz, G., Hernández, P., Schvartzman, J.B., Stasiak, A., and Krimer, D.B. (2009). Interplay of DNA supercoiling and catenation during the segregation of sister duplexes. Nucleic Acids Res. *37*, 5126–5137.

Matsumoto, K., and Hirose, S. (2004). Visualization of unconstrained negative supercoils of DNA on polytene chromosomes of Drosophila. J. Cell. Sci. *117*, 3797–3805.

McClendon, A.K., Rodriguez, A.C., and Osheroff, N. (2005). Human topoisomerase IIalpha rapidly relaxes positively supercoiled DNA: implications for enzyme action ahead of replication forks. J. Biol. Chem. *280*, 39337–39345.

McNamara, S., Wang, H., Hanna, N., and Miller, W.H., Jr (2008). Topoisomerase IIbeta negatively modulates retinoic acid receptor alpha function: a novel mechanism of retinoic acid resistance. Mol. Cell. Biol. *28*, 2066–2077.

Miassod, R., Razin, S.V., and Hancock, R. (1997). Distribution of topoisomerase II-mediated cleavage sites and relation to structural and functional landmarks in 830 kb of Drosophila DNA. Nucleic Acids Res. *25*, 2041–2046.

Mitsui, J., Takahashi, Y., Goto, J., Tomiyama, H., Ishikawa, S., Yoshino, H., Minami, N., Smith, D.I., Lesage, S., Aburatani, H., et al. (2010). Mechanisms of genomic instabilities underlying two common fragile-site-associated loci, PARK2 and DMD, in germ cell and cancer cell lines. Am. J. Hum. Genet. *87*, 75–89.

Mizutani, M., Ura, K., and Hirose, S. (1991a). DNA superhelicity affects the formation of transcription preinitiation complex on eukaryotic genes differently. Nucleic Acids Res. *19*, 2907–2911.

Mizutani, M., Ohta, T., Watanabe, H., Handa, H., and Hirose, S. (1991b). Negative supercoiling of DNA facilitates an interaction between transcription factor IID and the fibroin gene promoter. Proc. Natl. Acad. Sci. U.S.A. *88*, 718–722.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., De Laat, W., Spitz, F., and Duboule, D. (2011). A regulatory archipelago controls Hox genes transcription in digits. Cell *147*, 1132–1145.

Morham, S.G., Kluckman, K.D., Voulomanos, N., and Smithies, O. (1996). Targeted disruption of the mouse topoisomerase I gene by camptothecin selection. Mol. Cell. Biol. *16*, 6804–6809.

Morse, R.H., Pederson, D.S., Dean, A., and Simpson, R.T. (1987). Yeast nucleosomes allow thermal untwisting of DNA. Nucleic Acids Res. *15*, 10311–10330.

Naughton, C., Avlonitis, N., Corless, S., Prendergast, J.G., Mati, I.K., Eijk, P.P., Cockroft, S.L., Bradley, M., Ylstra, B., and Gilbert, N. (2013a). Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nat. Struct. Mol. Biol. *20*, 387–395.

Naughton, C., Corless, S., and Gilbert, N. (2013b). Divergent RNA transcription: A role in promoter unwinding? Transcription *4*.

Neidhardt, F.C., and Curtiss, R. (1999). Escherichia coli and Salmonella cellular and molecular biology (Washington, D.C.: ASM Press).

Németh, A., and Längst, G. (2004). Chromatin higher order structure: opening up chromatin for transcription. Brief Funct Genomic Proteomic *2*, 334–343.

Ng, H.L., Kopka, M.L., and Dickerson, R.E. (2000). The structure of a stable intermediate in the A <--> B DNA helix transition. Proc. Natl. Acad. Sci. U.S.A. *97*, 2035–2039.

Nitiss, J.L. (2009a). DNA topoisomerase II and its growing repertoire of biological functions. Nat. Rev. Cancer *9*, 327–337.

Nitiss, J.L. (2009b). Targeting DNA topoisomerase II in cancer chemotherapy. Nat. Rev. Cancer *9*, 338–350.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Nordheim, A., Pardue, M.L., Lafer, E.M., Möller, A., Stollar, B.D., and Rich, A. (1981). Antibodies to left-handed Z-DNA bind to interband regions of Drosophila polytene chromosomes. Nature *294*, 417–422.

O'Neill, T.E., Meersseman, G., Pennings, S., and Bradbury, E.M. (1995). Deposition of histone H1 onto reconstituted nucleosome arrays inhibits both initiation and elongation of transcripts by T7 RNA polymerase. Nucleic Acids Res. *23*, 1075–1082.

Patikoglou, G., and Burley, S.K. (1997). Eukaryotic transcription factor-DNA complexes. Annu Rev Biophys Biomol Struct *26*, 289–325.

Paulson, J.R., and Laemmli, U.K. (1977). The structure of histone-depleted metaphase chromosomes. Cell *12*, 817–828.

Pennings, S., Meersseman, G., and Bradbury, E.M. (1994). Linker histones H1 and H5 prevent the mobility of positioned nucleosomes. Proc. Natl. Acad. Sci. U.S.A. *91*, 10275–10279.

Peter, B.J., Arsuaga, J., Breier, A.M., Khodursky, A.B., Brown, P.O., and Cozzarelli, N.R. (2004). Genomic transcriptional response to loss of chromosomal supercoiling in Escherichia coli. Genome Biol. *5*, R87.

Peters, A.H., O'Carroll, D., Scherthan, H., Mechtler, K., Sauer, S., Schöfer, C., Weipoltshammer, K., Pagani, M., Lachner, M., Kohlmaier, A., et al. (2001). Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. Cell *107*, 323–337.

Pettijohn, D.E., and Pfenninger, O. (1980). Supercoils in prokaryotic DNA restrained in vivo. Proc. Natl. Acad. Sci. U.S.A. *77*, 1331–1335.

Pinheiro, I., Margueron, R., Shukeir, N., Eisold, M., Fritzsch, C., Richter, F.M., Mittler, G., Genoud, C., Goyama, S., Kurokawa, M., et al. (2012). Prdm3 and Prdm16 are H3K9me1 methyltransferases required for mammalian heterochromatin integrity. Cell *150*, 948–960.

Pommier, Y. (2006). Topoisomerase I inhibitors: camptothecins and beyond. Nat. Rev. Cancer *6*, 789–802.

Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. *38*, D105–110.

Postow, L., Crisona, N.J., Peter, B.J., Hardy, C.D., and Cozzarelli, N.R. (2001). Topological challenges to DNA replication: conformations at the fork. Proc. Natl. Acad. Sci. U.S.A. *98*, 8219–8226.

Postow, L., Hardy, C.D., Arsuaga, J., and Cozzarelli, N.R. (2004). Topological domain structure of the Escherichia coli chromosome. Genes Dev. *18*, 1766–1779.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851–1854.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rich, A., and Zhang, S. (2003). Timeline: Z-DNA: the long road to biological function. Nat. Rev. Genet. *4*, 566–572.

Richmond, T.J., and Davey, C.A. (2003). The structure of DNA in the nucleosome core. Nature *423*, 145–150.

Roca, J., Ishida, R., Berger, J.M., Andoh, T., and Wang, J.C. (1994). Antitumor bisdioxopiperazines inhibit yeast DNA topoisomerase II by trapping the enzyme in the form of a closed protein clamp. Proc. Natl. Acad. Sci. U.S.A. *91*, 1781–1785.

Rogakou, E.P., Pilch, D.R., Orr, A.H., Ivanova, V.S., and Bonner, W.M. (1998). DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. J. Biol. Chem. *273*, 5858–5868.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annu. Rev. Biochem. *79*, 233–269.

Saffran, W.A., Welsh, J.T., Knobler, R.M., Gasparro, F.P., Cantor, C.R., and Edelson, R.L. (1988). Preparation and characterization of biotinylated psoralen. Nucleic Acids Res. *16*, 7221–7231.

Saha, A., Wittmeyer, J., and Cairns, B.R. (2006). Chromatin remodelling: the industrial revolution of DNA around histones. Nat. Rev. Mol. Cell Biol. *7*, 437–447.

Salceda, J., Fernández, X., and Roca, J. (2006). Topoisomerase II, not topoisomerase I, is the proficient relaxase of nucleosomal DNA. EMBO J. *25*, 2575–2583.

Sano, K., Miyaji-Yamaguchi, M., Tsutsui, K.M., and Tsutsui, K. (2008). Topoisomerase IIbeta activates a subset of neuronal genes that are repressed in AT-rich genomic environment. PLoS ONE *3*, e4103.

Sarma, K., and Reinberg, D. (2005). Histone variants meet their match. Nat. Rev. Mol. Cell Biol. *6*, 139–149.

Schalch, T., Duda, S., Sargent, D.F., and Richmond, T.J. (2005). X-ray structure of a tetranucleosome and its implications for the chromatin fibre. Nature *436*, 138–141.

Schoeffler, A.J., and Berger, J.M. (2008). DNA topoisomerases: harnessing and constraining energy to govern chromosome topology. Q. Rev. Biophys. *41*, 41–101.

Schwartz, M., Zlotorynski, E., Goldberg, M., Ozeri, E., Rahat, A., Le Sage, C., Chen, B.P.C., Chen, D.J., Agami, R., and Kerem, B. (2005). Homologous recombination and nonhomologous end-joining repair pathways regulate fragile site stability. Genes Dev. *19*, 2715–2726.

Schwartz, T., Behlke, J., Lowenhaupt, K., Heinemann, U., and Rich, A. (2001). Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins. Nat. Struct. Biol. *8*, 761–765.

Scott, D.J., Harding, S.E., and Rowe, A.J. (2005). Analytical ultracentrifugation techniques and methods (Cambridge: RSC Publishing).

Sedat, J., and Manuelidis, L. (1978). A direct approach to the structure of eukaryotic chromosomes. Cold Spring Harb. Symp. Quant. Biol. *42 Pt 1*, 331–350.

Segal, E., and Widom, J. (2009). Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr. Opin. Struct. Biol. *19*, 65–71.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Seila, A.C., Core, L.J., Lis, J.T., and Sharp, P.A. (2009). Divergent transcription: a new feature of active promoters. Cell Cycle *8*, 2557–2564.

She, X., Rohl, C.A., Castle, J.C., Kulkarni, A.V., Johnson, J.M., and Chen, R. (2009). Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. BMC Genomics *10*, 269.

Sherratt, D.J. (2003). Bacterial chromosome dynamics. Science *301*, 780–785.

Shlyakhtenko, L.S., Hsieh, P., Grigoriev, M., Potaman, V.N., Sinden, R.R., and Lyubchenko, Y.L. (2000). A cruciform structural transition provides a molecular switch for chromosome structure and dynamics. J. Mol. Biol. *296*, 1169–1173.

Shogren-Knaak, M., Ishii, H., Sun, J.-M., Pazin, M.J., Davie, J.R., and Peterson, C.L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science *311*, 844–847.

Sims, R.J., 3rd, Chen, C.-F., Santos-Rosa, H., Kouzarides, T., Patel, S.S., and Reinberg, D. (2005). Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. J. Biol. Chem. *280*, 41789–41792.

Sinden, R.R. (1994). DNA structure and function (San Diego: Academic Press).

Sinden, R.R., and Pettijohn, D.E. (1981). Chromosomes in living Escherichia coli cells are segregated into domains of supercoiling. Proc. Natl. Acad. Sci. U.S.A. *78*, 224–228.

Sinden, R.R., Carlson, J.O., and Pettijohn, D.E. (1980). Torsional tension in the DNA double helix measured with trimethylpsoralen in living E. coli cells: analogous measurements in insect and human cells. Cell *21*, 773–783.

Singleton, M.R., Dillingham, M.S., and Wigley, D.B. (2007). Structure and mechanism of helicases and nucleic acid translocases. Annu. Rev. Biochem. *76*, 23–50.

Song, P.S., and Ou, C.N. (1980). Labeling of nucleic acids with psoralens. Ann. N. Y. Acad. Sci. *346*, 355–367.

Srivenugopal, K.S., Lockshon, D., and Morris, D.R. (1984). Escherichia coli DNA topoisomerase III: purification and characterization of a new type I enzyme. Biochemistry *23*, 1899–1906.

Staynov, D.Z. (2008). The controversial 30 nm chromatin fibre. Bioessays *30*, 1003–1009.

Stewart, A.F., Herrera, R.E., and Nordheim, A. (1990). Rapid induction of c-fos transcription reveals quantitative linkage of RNA polymerase II and DNA topoisomerase I enzyme activities. Cell *60*, 141–149.

Stewart, L., Redinbo, M.R., Qiu, X., Hol, W.G., and Champoux, J.J. (1998). A model for the mechanism of human topoisomerase I. Science *279*, 1534–1541.

Struhl, K., and Segal, E. (2013). Determinants of nucleosome positioning. Nat. Struct. Mol. Biol. *20*, 267–273.

Tabuchi, H., and Hirose, S. (1988). DNA supercoiling facilitates formation of the transcription initiation complex on the fibroin gene promoter. J. Biol. Chem. *263*, 15282–15287.

Le Tallec, B., Dutrillaux, B., Lachages, A.-M., Millot, G.A., Brison, O., and Debatisse, M. (2011). Molecular profiling of common fragile sites in human fibroblasts. Nat. Struct. Mol. Biol. *18*, 1421–1423.

Le Tallec, B., Millot, G.A., Blin, M.E., Brison, O., Dutrillaux, B., and Debatisse, M. (2013). Common Fragile Site Profiling in Epithelial and Erythroid Cells Reveals that Most Recurrent Cancer Deletions Lie in Fragile Sites Hosting Large Genes. Cell Rep.

Taniguchi, T., and Takayama, S. (1986). High-order structure of metaphase chromosomes: evidence for a multiple coiling model. Chromosoma *93*, 511–514.

Teves, S.S., and Henikoff, S. (2014). Transcription-generated torsional stress destabilizes nucleosomes. Nat. Struct. Mol. Biol. *21*, 88–94.

Thamann, T.J., Lord, R.C., Wang, A.H., and Rich, A. (1981). The high salt form of poly(dG-dC).poly(dG-dC) is left-handed Z-DNA: Raman spectra of crystals and solutions. Nucleic Acids Res. *9*, 5443–5457.

Thoma, F., and Koller, T. (1977). Influence of histone H1 on chromatin structure. Cell *12*, 101–107.

Thoma, F., Koller, T., and Klug, A. (1979). Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. J. Cell Biol. *83*, 403–427.

Thomas, J.O. (1999). Histone H1: location and role. Curr. Opin. Cell Biol. *11*, 312–317.

Toedling, J. (2012). Ringo - R Investigation of NimbleGen Oligoarrays.

Toedling, J., Skylar, O., Sklyar, O., Krueger, T., Fischer, J.J., Sperling, S., and Huber, W. (2007). Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts. BMC Bioinformatics *8*, 221.

Tolstorukov, M.Y., Goldman, J.A., Gilbert, C., Ogryzko, V., Kingston, R.E., and Park, P.J. (2012). Histone variant H2A.Bbd is associated with active transcription and mRNA processing in human cells. Mol. Cell *47*, 596–607.

Tomicic, M.T., and Kaina, B. (2013). Topoisomerase degradation, DSB repair, p53 and IAPs in cancer cell resistance to camptothecin-like topoisomerase I inhibitors. Biochim. Biophys. Acta *1835*, 11–27.

Udvardy, A., and Schedl, P. (1991). Chromatin structure, not DNA sequence specificity, is the primary determinant of topoisomerase II sites of action in vivo. Mol. Cell. Biol. *11*, 4973–4984.

Varga-Weisz, P.D., Wilm, M., Bonte, E., Dumas, K., Mann, M., and Becker, P.B. (1997). Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. Nature *388*, 598–602.

Villeponteau, B., Lundell, M., and Martinson, H. (1984). Torsional stress promotes the DNAase I sensitivity of active genes. Cell *39*, 469–478.

Vinograd, J., Lebowitz, J., Radloff, R., Watson, R., and Laipis, P. (1965). The twisted circular form of polyoma viral DNA. Proc. Natl. Acad. Sci. U.S.A. *53*, 1104–1111.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Wang, A.H., Quigley, G.J., Kolpak, F.J., Crawford, J.L., Van Boom, J.H., Van der Marel, G., and Rich, A. (1979). Molecular structure of a left-handed double helical DNA fragment at atomic resolution. Nature *282*, 680–686.

Wang, J.C., Peck, L.J., and Becherer, K. (1983). DNA supercoiling and its effects on DNA structure and function. Cold Spring Harb. Symp. Quant. Biol. *47 Pt 1*, 85–91.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. *39*, 457–466.

Weintraub, H., Cheng, P.F., and Conrad, K. (1986). Expression of transfected DNA depends on DNA topology. Cell *46*, 115–122.

De Wit, E., and De Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. Genes Dev. *26*, 11–24.

De Wit, E., Bouwman, B.A.M., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J.A.M., Krijger, P.H.L., Festuccia, N., Nora, E.P., Welling, M., et al. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature.

Wittig, B., Dorbic, T., and Rich, A. (1989). The level of Z-DNA in metabolically active, permeabilized mammalian cell nuclei is regulated by torsional strain. J. Cell Biol. *108*, 755–764.

Wittig, B., Dorbic, T., and Rich, A. (1991). Transcription is associated with Z-DNA formation in metabolically active permeabilized mammalian cell nuclei. Proc. Natl. Acad. Sci. U.S.A. *88*, 2259–2263.

Wittig, B., Wölfl, S., Dorbic, T., Vahrson, W., and Rich, A. (1992). Transcription of human c-myc in permeabilized nuclei is associated with formation of Z-DNA in three discrete regions of the gene. EMBO J. *11*, 4653–4663.

Woessner, R.D., Mattern, M.R., Mirabelli, C.K., Johnson, R.K., and Drake, F.H. (1991). Proliferation- and cell cycle-dependent differences in expression of the 170 kilodalton and 180 kilodalton forms of topoisomerase II in NIH-3T3 cells. Cell Growth Differ. *2*, 209–214.

Wojciechowska, M., Napierala, M., Larson, J.E., and Wells, R.D. (2006). Non-B DNA conformations formed by long repeating tracts of myotonic dystrophy type 1, myotonic dystrophy type 2, and Friedreich's ataxia genes, not the sequences per se, promote mutagenesis in flanking regions. J. Biol. Chem. *281*, 24531–24543.

Wolffe, A. (1998). Chromatin : structure and function (San Diego: Academic Press).

Worcel, A., and Burgi, E. (1972). On the structure of the folded chromosome of Escherichia coli. J. Mol. Biol. *71*, 127–147.

Wu, L., and Hickson, I.D. (2003). The Bloom's syndrome helicase suppresses crossing over during homologous recombination. Nature *426*, 870–874.

Wu, C.-C., Li, T.-K., Farh, L., Lin, L.-Y., Lin, T.-S., Yu, Y.-J., Yen, T.-J., Chiang, C.-W., and Chan, N.-L. (2011). Structural basis of type II topoisomerase inhibition by the anticancer drug etoposide. Science *333*, 459–462.

Wu, H., Caffo, B., Jaffee, H.A., Irizarry, R.A., and Feinberg, A.P. (2010). Redefining CpG islands using hidden Markov models. Biostatistics *11*, 499–514.

Wysocka, J., Swigut, T., Xiao, H., Milne, T.A., Kwon, S.Y., Landry, J., Kauer, M., Tackett, A.J., Chait, B.T., Badenhorst, P., et al. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature *442*, 86–90.

Yager, T.D., McMurray, C.T., and Van Holde, K.E. (1989). Salt-induced release of DNA from nucleosome core particles. Biochemistry *28*, 2271–2281.

Yang, J., Bachrati, C.Z., Ou, J., Hickson, I.D., and Brown, G.W. (2010). Human topoisomerase IIIalpha is a single-stranded DNA decatenase that is stimulated by BLM and RMI1. J. Biol. Chem. *285*, 21426–21436.

Zechiedrich, E.L., Khodursky, A.B., and Cozzarelli, N.R. (1997). Topoisomerase IV, not gyrase, decatenates products of site-specific recombination in Escherichia coli. Genes Dev. *11*, 2580–2592.

Zhang, C.X., Chen, A.D., Gettel, N.J., and Hsieh, T.S. (2000). Essential functions of DNA topoisomerase I in Drosophila melanogaster. Dev. Biol. *222*, 27–40.

Zhang, Z., Wippo, C.J., Wal, M., Ward, E., Korber, P., and Pugh, B.F. (2011). A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. Science *332*, 977–980.

Zlotorynski, E., Rahat, A., Skaug, J., Ben-Porat, N., Ozeri, E., Hershberg, R., Levi, A., Scherer, S.W., Margalit, H., and Kerem, B. (2003). Molecular basis for expression of common and rare fragile sites. Mol. Cell. Biol. *23*, 7143–7151.

(1994). Chromosome analysis protocols (Totowa, N.J: Humana Press).